Relation Extraction in Underexplored Biomedical Domains: A Diversity-optimized Sampling and Synthetic Data Generation Approach

Maxime Delmas*
Idiap Research Institute
Switzerland
maxime.delmas@idiap.ch

Magdalena Wysocka
Digital Cancer Research,
CRUK National Biomarker Centre
United Kingdom
magdalena.wysocka@digitalecmt.org

André Freitas
Idiap Research Institute
Switzerland
Department of Computer Science,
University of Manchester
Digital Cancer Research,
CRUK National Biomarker Centre
United Kingdom
andre.freitas@idiap.ch,
andre.freitas@manchester.ac.uk

The sparsity of labeled data is an obstacle to the development of Relation Extraction (RE) models and the completion of databases in various biomedical areas. While being of high interest in drug-discovery, the literature on natural products, reporting the identification of potential bioactive compounds from organisms, is a concrete example of such an overlooked topic. To mark the start of this new task, we created the first curated evaluation dataset and extracted literature items from the LOTUS database to build training sets. To this end, we developed a new sampler, inspired by diversity metrics in ecology, named Greedy Maximum Entropy sampler (`https://github.com/idiap/gme-sampler`). The strategic optimization of both balance and

* Corresponding author.

Action Editor: Byron Wallace. Submission received: 23 November 2023; revised version received: 9 March 2024; accepted for publication: 19 April 2024.

https://doi.org/10.1162/coli_a_00520

diversity of the selected items in the evaluation set is important given the resource-intensive nature of manual curation. After quantifying the noise in the training set, in the form of discrepancies between the text of input abstracts and the expected output labels, we explored different strategies accordingly. Framing the task as an end-to-end Relation Extraction, we evaluated the performance of standard fine-tuning (BioGPT, GPT-2, and Seq2rel) and few-shot learning with open Large Language Models (LLMs) (LLaMA 7B-65B). In addition to their evaluation in few-shot settings, we explore the potential of open LLMs as synthetic data generators and propose a new workflow for this purpose. All evaluated models exhibited substantial improvements when fine-tuned on synthetic abstracts rather than the original noisy data. We provide our best performing (F1-score = 59.0) BioGPT-Large model for end-to-end RE of natural products relationships along with all the training and evaluation datasets. See more details at `https://github.com/idiap/abroad-re`.

1. Introduction

The biomedical literature constitutes a vast but still underexploited reservoir of knowledge, the growth of which reflects the expansion of topics and areas of applications. However, the diversity and morphological richness of bio-entities and the complexity of the relationships expressed between them contrast with the sparsity of the available labeled data. While some domains can already benefit from efficient extraction models (e.g., chemical–disease relationships) for database completion, less popular domains, like the literature on natural products (NP), are often overlooked. NPs are chemical compounds produced by living organisms (plants, bacteria, fungi, etc.) exhibiting a wide range of structure and functions and offering a vast reservoir of potential therapeutic molecules. The isolation and identification of NP is primarily reported in the scientific literature and also disseminated in different public databases (e.g., COCONUT Sorokina et al. 2021; KNApSAcK Shinbo et al. 2006). Recently, the LOTUS initiative (Rutz et al. 2022) has successfully established an Open and FAIR standard resource for natural products chemistry through a rigorous harmonization of a heterogenous set of databases. However, the extent of the NP landscape is not reflected by the content of the databases, which are incomplete and exhibit an imbalanced coverage toward model organisms (e.g., *A. Thaliana*). While a significant portion of the existing literature remains unannotated, there is also a continuous surge of new publications reporting novel relationships that could contribute to filling this gap.

Enriching such knowledge bases requires jointly performing Named Entity Recognition (NER) and Relation Extraction (RE). In this case, NER is defined as a sub-task that consists of identifying the boundaries and classifying the type of named entities (i.e., an organism "*Isaria sinclairii*" and a chemical "fingolimod"[1]). The subsequent RE step is the semantic classification of the relations between two (or more) entities. To complete NP databases, the objective is to extract the "*produces*" or "*is isolated from*" relationships between organisms and chemicals. Note that other types of relationships can also be expressed, such as "*inhibits the growth of*". Traditional deep learning models exhibiting SOTA performance on NER and RE (separately or in so-called *end-to-end* models) rely on a large set of labeled data (Luo et al. 2022b; Giorgi, Bader, and Wang 2022; Wang et al. 2020). However, while datasets like Linneaus (Gerner, Nenadic, and Bergman 2010) have been successfully applied for organism recognition,

1 `https://lotus.naturalproducts.net/compound/lotus_id/LTS0203935`.

existing chemical NER datasets, that is, CHEMDNER (Krallinger et al. 2015), do not provide sufficient coverage on the NP literature and do not adequately capture their morphological specificities. Along with the typically long systematic names of metabolites (e.g., `3'-[gamma-hydroxymethyl-(E)-gamma-methylallyl]-2,4,2',4' -tetrahydroxychalcone 11'-O-coumarate`[2]), many chemical mentions are defined as multiple co-joined enumerations, where entities are mentioned in non-continuous strings such as "cystodiones A-D" or "wortmannins C and D", and are particularly frequent. These chemical mentions must be correctly identified and expanded to recover the full list of entities, which also adds complexity to the decoding process. Finally, to the best of our knowledge, no datasets are available for the subsequent RE step (Luo et al. 2022a). The aforementioned constraints are frequently encountered in BioNLP, when venturing beyond the well-studied chemical–disease associations or protein–protein interactions.

Meanwhile, the abundance of unlabeled textual data has been instrumental in driving recent breakthroughs in representation learning (Wysocki et al. 2023) and the development of the foundational models (e.g., GPT and LLaMA model families). The zero/few-shot learning capabilities of Large Language Models (LLMs) (Kojima et al. 2023; Brown et al. 2020) make them serious candidates for performing a task with only a handful of examples. Moreover, conversation (chatbots) and instruction-tuned models (Zhang et al. 2023) also represent a promising opportunity for synthetic data generation to alleviate the main problem, namely, the lack of labeled data within the target distribution. Indeed, beyond the sophistication of model architectures, data availability and quality are limiting factors for the extraction performance, but often neglected (Sambasivan et al. 2021).

In order to address these scarcity constraints, we propose an end-to-end generative extraction paradigm, which introduces two novel methodological contributions. Firstly, we introduce a diversity-optimized sampling strategy, which minimizes the selection of items for the parsimonious creation of evaluation gold-standards and training sets. This component minimizes the popularity biases and associated imbalance towards entities which are over-expressed in the literature (e.g., model organisms and recurring substances), allowing for a systematic (entropy-based) method to maximize diversity and measure the utility of new annotations. Secondly, we use the generative expressivity of models fine-tuned on conversations and instructions for creating within-distribution synthetic data, to support the construction of end-to-end joint NER-RE extraction models. In this framework, the diversity-sampled entities and associated relations are linguistically embedded within synthetically generated text. The overall framework is depicted in Figure 1.

More formally, this article aims to investigate the following research hypotheses (RHs) as supporting mechanisms for addressing these limitations, using NP as a validation domain:

- **RH1: Diversity-optimized sampling provides a valuable selection of items to build training and evaluation datasets for RE.**
- **RH2: In a practical scenario with noisy labels, LLMs can be more beneficial as a synthetic data generator than unsupervised predictors.**

2 `https://pubmed.ncbi.nlm.nih.gov/11678652`.

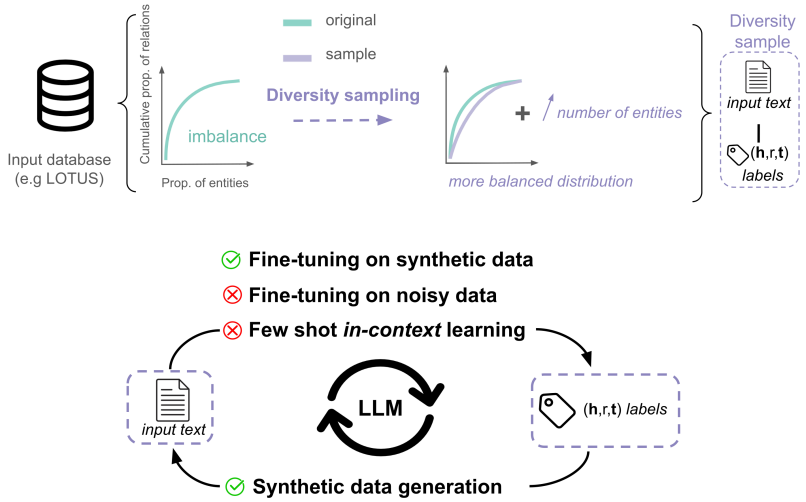


Figure 1
Combining diversity sampling and synthetic data generation. Upper part: With unbalanced data, annotated relationships are frequently associated with the same head or tail entities. The proposed sampling method is used to build training and evaluation datasets by maximizing the diversity of entities in the extracted samples. The larger the set of entities and the more they are uniformly distributed across the relations, the higher the diversity. Bottom part: When labeled data are scarce or noisy, LLMs prove more advantageous in the reverse task, as synthetic data generators, rather than as few-shot learners. Subsequent models trained on synthetic data outperform few-shot learning and models trained on raw noisy data.

1.1 Related Work

Biomedical RE (Shahab 2017; Zhao et al. 2020, 2023) encompasses various subtypes, depending on the considered bio-entities, such as drug–drug interactions (Zhang, Leng, and Liu 2020), chemical–disease relationships (Li et al. 2016), gene–disease associations (Su et al. 2021), and protein–protein interactions (Ahmed et al. 2019), among the most popular. Investigating the overlooked NP relationships necessitated the exploration of several interconnected sub-tasks, including the selection and partitioning of a dataset, the generation of synthetic data, and the assessment of various end-to-end RE strategies. This section provides a review of the closely related works that align with these three development axes.

1.1.1 Splitting Datasets and Impact of Diversity. Data selection and partitioning methods can significantly impact the generalization performance of supervised models. Xu and Goodacre (2018) evaluated various splitting techniques, including K-S (Kennard and Stone 1969) and SPXY (Galvao et al. 2005), and emphasized the importance of maintaining a balance between training and test sets for a reliable evaluation of models. Like the recently proposed SPlit (Joseph and Vakayil 2022) method, these approaches aim to select a representative subset of the data, leveraging different distance metrics. Unlike the Euclidean or energy-based distances used in aforementioned methods, the Greedy Maximum Entropy (GME)-sampler uses an entropy-based metric to capture diversity and select representative evaluation and training sets. Although these distance-based

methods share a common objective, they were initially designed to work with continuous variables, rather than categorical variables, such as large sets of organisms and chemicals. Moreover, to the best of our knowledge, no method has been specifically developed to sample documents reporting `N-ary` relations for the purpose of building NER/RE datasets. The GME-sampler represents a first attempt to address this gap. Regarding diversity, Yu, Khadivi, and Xu (2022) investigated various diversity-based metrics for selecting training data, and demonstrated their positive impact on the performance of NER models. Additionally, other works have highlighted the significance of effective data selection over a naive increase of the dataset size for training (Axelrod, He, and Gao 2011; Fan et al. 2017; Feng et al. 2018).

1.1.2 Synthetic Data Generation. Training neural RE models strongly rely on a substantial and diverse set of training data. However, annotating large datasets with experts is time-consuming and costly. To overcome this limitation, many studies explored approaches such as Data Augmentation (DA) (Hu et al. 2023; Feng et al. 2021; Pellicer, Ferreira, and Costa 2023) and Distant Supervision (DS) (Smirnova and Cudré-Mauroux 2018; Mintz et al. 2009), which enable the expansion of the dataset size by creating new training examples from existing ones, or, by assigning pseudo-labels to external, unlabeled data. In the biomedical domain, the RE challenge ChemProt (Yoon et al. 2023; Iinuma, Miwa, and Sasaki 2022) or protein–protein interactions extraction (Su et al. 2019), have recently benefited from the application of these methods. Synthetic data generation (SDG) goes beyond DA or DS by creating fully synthetic datasets, namely, paired input text and output labels. A significant body of influential works has leveraged the generative capabilities of LLMs to propose different SDG strategies in zero-shot (Ye et al. 2022; Gao et al. 2023; Schick and Schütze 2021; He et al. 2022; Wang et al. 2021; Smith et al. 2024; Meng et al. 2022; Kumar, Choudhary, and Cho 2020), few-shot (Bonifacio et al. 2022; Dai et al. 2023; Meng et al. 2023; Chen et al. 2022a; Yoo et al. 2021), or by fine-tuning (Anaby-Tavor et al. 2020; Papanikolaou and Pierleoni 2020; Hartvigsen et al. 2022). Similarly to this work, Josifoski et al. (2023) also proposed to reverse the task and used LLMs from OpenAI to generate plausible input text based on expected output triplets from Wikidata. Tang et al. (2023) compared the performance of an LLM (ChatGPT) in directly extracting information from unstructured clinical text to its potential use as synthetic data generator for DA. Veselovsky et al. (2023) evaluated various prompting strategies to improve diversity and alignment between synthetic and real-world data distributions for sarcasm detection. Yang et al. (2020) combined synthetic data generation with a diversity-augmentation component for common sense reasoning. Aggarwal, Jin, and Ahmad (2023) applied SDG to biomedical NER, while Xu et al. (2023) used a two-stage training procedure on synthetic and golden data, notably for extracting protein interactions with the ChemProt dataset. In contrast, this work proposes to leverage Open LLMs to generate synthetic abstracts based on a list of verbalized main findings. The diversity of the generations is increased and guided by the entropy-based sampling of the seed articles which originally report these findings, as well as a set of crafted patterns of expressions.

1.1.3 End-to-End Relation Extraction. Kambar, Esmaeilzadeh, and Heidari (2022) classifies various strategies and highlights the potential of end-to-end (or joint) NER and RE methods to overcome limitations of the traditional pipeline approaches. In the biomedical domain, Li et al. (2017) proposed a Bi-LSTM for drug adverse effects extraction, while Esmail Zadeh Nojoo Kambar, Esmaeilzadeh, and Taghva (2022) introduced a graph neural network for chemical–protein interactions. Recent approaches frame the

task in a generative "text-to-text" process, using sequence-to-sequence models, by linearalizing the expected relations as a text string to be decoded from the input. Seq2Rel (Giorgi, Bader, and Wang 2022) and REBEL (Huguet Cabot and Navigli 2021) proposed different linearization schemas, and Zhang et al. (2020) and Zeng et al. (2019) notably assessed the biases caused by the forced order of relationships during training. Hou et al. (2022) trained a sequence-to-sequence model for drug–target interactions extraction, and Zeng et al. (2018) introduced a copy mechanism. Additionally, Eberts and Ulges (2021) used four task-specific sub-components, and Paolini et al. (2021) utilized a translation mechanism. Finally, BioGPT (Luo et al. 2022b) demonstrated SOTA performance on several biomedical datasets using an autoregressive approach, providing the input text as context.

With the aim of providing an end-to-end RE model to help expanding NP databases, we started by building a training and evaluation dataset. Inspired by the metrics used in ecology, we first proposed the Greedy-Maximum-Entropy sampler (GME-sampler) to extract a diversity-optimized sample from the LOTUS database. By manually annotating the top-diverse items, we proposed the first evaluation dataset for this task, which can serve as a benchmark for future developments in this area. Following a descriptive analysis of the remaining data and quantifying the noise present in the form of discrepancies between raw input text and annotated (standardized) labels, we evaluated various modeling approaches. First, we compared the performance of standard fine-tuning techniques on the available noisy data to the few-shot learning capabilities of open LLMs. Leveraging the generative capabilities of a LLM (Vicuna-13B), we then proposed a novel synthetic abstract generation pipeline and demonstrated the significant performance improvements (on average 24.7% in F1-score) brought by these new training data on the evaluated models. In line with these results, we have made available our best-performing BioGPT-Large model (F1-score $= 59.0$) and the $\approx$ 25,000 synthetic abstracts on which it has been trained. A synthetic diagram of the different strategies explored in this work is presented in Figure 2. The main contributions of the work can be summarized as:

- A diversity-optimized sampler (GME-sampler) for building diverse and balanced datasets for RE (see `https://github.com/idiap/gme-sampler`).
- The first curated evaluation dataset for RE between organisms and NP (see `https://zenodo.org/records/8422007`).
- An evaluation of different strategies for RE with noisy labels.
- A framework for synthetic data generation via chatbot or instruction-tuned models and the produced training datasets (see `https://github.com/idiap/abroad-re` and `https://zenodo.org/records/8422294`).
- A set of ready-to-use BioGPT fine-tuned models (see `https://huggingface.co/mdelmas/BioGPT-Large-Natural-Products-RE-Diversity-synt-v1.0`)

2. Proposed Approach

This section describes the different methodology used in this work. We start by describing our first contribution, the GME-sampler, in Section 2.1. The few-shot learning and

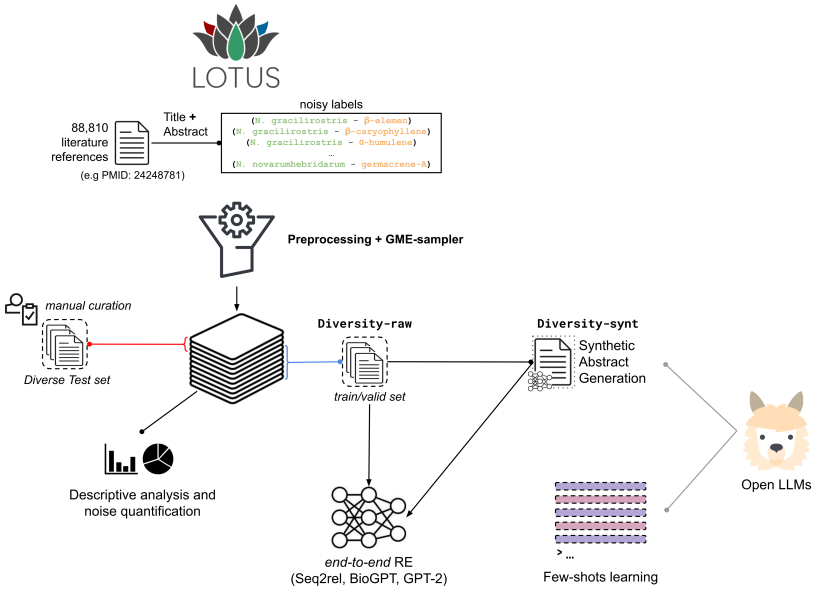


Figure 2
A global diagram of the workflow presented in this work.

fine-tuning strategies evaluated for the RE task are then described in Sections 2.2.1 and 2.2.2. The synthetic abstract generation procedure is described in 2.3. Finally, details on the evaluation, experimental setup and implementation details are provided in Appendix A.

2.1 Greedy Maximum Entropy Sampling (GME)

The objective is to extract a sample S of documents from an initial set D with an optimized diversity of mentioned organisms and chemicals: $S \subset D$, $|S| = l$ and $|D| = L$. The initial set D corresponds to the LOTUS dataset, in which each document d reports a set of relations between organism(s) and isolated natural product(s): $d = \{r_1, r_2, \ldots, r_{n_d}\}$, where n_d is the number of reported relations in d. A relation $r_k = (o_i, c_j)$ involves the organism o_i and the chemical c_j. The set of organisms and chemicals are denoted as O and C, respectively.

Then, given a set S of documents, the probability that a reported relation involves the organism o_i (and similarly for the chemical c_j) is

$$P(o_i) = \frac{\sum_{d \in S} |\{r_k : o_i \in r_k\}^d|}{\sum_{d \in S} n_d}, \quad \begin{array}{l}\text{with } |\{r_k : o_i \in r_k\}^d| \\ \text{the number of relations involving} \\ o_i \text{ in } d\end{array} \tag{1}$$

It follows that the diversity of organisms or chemicals can be measured with the Shannon's entropy over the probability distributions of elements of O and C in the sample S. Expressed with entropy, the diversity reflects the uncertainty about the organism, or the chemical, which is attached to a relation reported in an article (Leinster 2021; Jost

2006). For organisms (respectively, chemicals) the Shannon's entropy in the sample S is $H_S(O) = -\sum_{o_i \in O} P(o_i) \log P(o_i)$. Adding a new document d to S will update the probability distributions of O (respectively, C), and the new observed entropy will be $H_{S_{+d}}(O)$ where $S_{+d} = S \cup \{d\}$. Therefore, to optimize the diversity over organisms and chemicals, the document $d*$ added to S minimizes the distance to the utopian point $(log|O|, log|C|)$ (maximal observable entropy over organisms and chemicals):

$$d* = \arg\min_{d} ||(H_{S_{+d}}(O), H_{S_{+d}}(C)) - (log|O|, log|C|)|| \quad (2)$$

The proposed sampling approach is a simple greedy algorithm that, at each step, selects and adds the new document $d*$ from D, maximizing the diversity on organisms and chemicals (see Equation 2). We refer to it as diversity-sampling. From an ecological perspective, a community's diversity can be regarded as low if a randomly chosen individual is more likely to be part of a *common* (dominant) species, and high if it is more probable to be part of a *rare* species (Hill 1973; Leinster 2021). Accordingly, the greater the number of species and the more uniformly distributed they are, the higher the diversity of the community. Shannon's entropy, which quantifies this uncertainty on the species, is used in ecology as a metric of diversity. In the context of an unbalanced dataset where annotated relationships are frequently associated with model organisms and recurring chemicals, this entropy-based criterion seems relevant to measure and select the documents that maximize the diversity during the sampling.

The method can also be seen as a ranking procedure, and a sample is determined by selecting the first top n ranked items. The selection of an appropriate sample size l is also a critical, but often overlooked factor. By monitoring $H_S(O)$ and $H_S(C)$ during the iterative construction of S (until $l = L$), it is possible to determine the step l at which diversity starts to deteriorate and sampling should be stopped, that is, when the new added documents provide relationships for already frequently reported entities in S. The GME-sampler, initially designed for the purpose of extracting data from LOTUS, has also been implemented as a standalone library. It is proposed as a method to build samples of documents reporting `N-ary` relations with optimized diversity, and can be applied in alternative contexts (e.g., Pharmacogenomics: *Variant–Drug–Adverse event*). See code available at `https://github.com/idiap/gme-sampler`.

2.2 Different Strategies for Relation Extraction

2.2.1 Few-shot In-context Learning with Open LLM. In few-shot settings, the model is prompted with K input–completion example pairs and one final input, with the objective of accurately generating the completion for the final input (Brown et al. 2020). Considering the limited size of the context-window (2,048 tokens), we carefully selected $K = 5$ archetypical parts of diverse abstracts that exemplify various patterns and specificities of reporting NP relationships. More details in Appendix A.1.

2.2.2 Fine-tuning. The task was framed as a special case of end-to-end Relation Extraction with a single relation. Several factors influenced this design: The need for a generalized NER component that is not dictionary-based, as new species and NP are discovered and named continuously, the scarcity of training data to segment the task in a NER-RE pipeline, and, the specific decoding requirements related to the NP relationships with multiple co-joined entities (e.g., gloeophyllins A-C). Given an input text X reporting relations $\{r_1, r_2, r_3\}$ between organisms o_1 and NP $c_{1:3}$ like: "Three new metabolites,

gloeophyllins A-C (1-3) have been isolated from the solid cultures of *Gloeophyllum abietinum*.", the expected output is the linearized list of relations Y: "*Gloeophyllum abietinum* produces gloeophyllin A; *Gloeophyllum abietinum* produces gloeophyllin B; *Gloeophyllum abietinum* produces gloeophyllin C", following recommendations from previous studies (Luo et al. 2022b; Giorgi, Bader, and Wang 2022). During fine-tuning, the objective function is then defined as:

$$L(\theta) = \sum_{t=1}^{|Y|} \log p(y_t | X, y_{<t}, \theta) \tag{3}$$

For efficient fine-tuning with minimal memory and parameter requirements, we adopted the QLoRA approach (Dettmers et al. 2023). The method extends the Low-rank Adaptation (LoRA) technique (Hu et al. 2022), which involves freezing the original model weights and training only a small set of parameters, known as adapters. Given the original weight of a layer W, with $h = Wx$, the adapters operate as an update to the initial weights $W + \Delta W$ and give $h = Wx + \Delta Wx$. With QLoRA, the LoRA strategy is integrated with quantization to minimize the memory footprint. It uses a low-precision storage data type (NF4) and a computation data type (BFloat16) for the forward and backward passes. Two models were evaluated using this strategy: BioGPT (Luo et al. 2022b) and GPT-2 *medium* (Radford et al. 2019). Both models share the same architecture, but BioGPT and its tokenizer were trained on PubMed items and achieved SOTA performance on three end-to-end RE tasks. As Luo et al. (2022b), we also add Seq2rel (Giorgi, Bader, and Wang 2022) as a second baseline using a sequence-to-sequence approach. More details on the choice of the models, experimental setup, hyperparameter tuning, and evaluation are presented in Appendices A.2–A.5.

2.3 Synthetic Abstract Generation

A general overview of the synthetic abstract generation is provided in Figure 3. The goal is to leverage the generative capabilities of instruction- and conversation-tuned models to correct the discrepancies between the expected output labels and the input text. Consistency is maintained by grounding the generated abstracts on key elements from an original seed abstract: title, keyphrases extracted from the abstract (and title), and verbalized main findings. The extracted keyphrases are also intended to mirror the annotated MeSH descriptors, which are attached to the title and abstract in a PubMed entry. The main findings represent the set of relations $\{r_1, r_2, \ldots, r_n\}$ between organisms and NP reported in the seed article according to the LOTUS database (the expected output labels). Both the keyphrase extraction and the subsequent generation step can be framed as instructions-guided tasks: "*Extract a list of keywords ...*", "*Create a scientific abstract ...*". As the extracted keywords and keyphrases will provide an essential context to constrain the generation of the synthetic abstracts, it is also arguably advantageous that both tasks are carried by the same model.

The extraction of keywords (illustrated in Box A of Figure 3) consists of prompting the model to extract keywords and keyphrases from the original abstract, and establish a coherent context for the subsequent generation. However, there is a risk that certain chemicals or organisms mentioned in the original abstract may also be extracted as keywords. They could be erroneously mixed by the model with the main findings (the expected output labels) in the generation step. This could result in the generation of abstracts with unintended relationships that were not specified in the original main

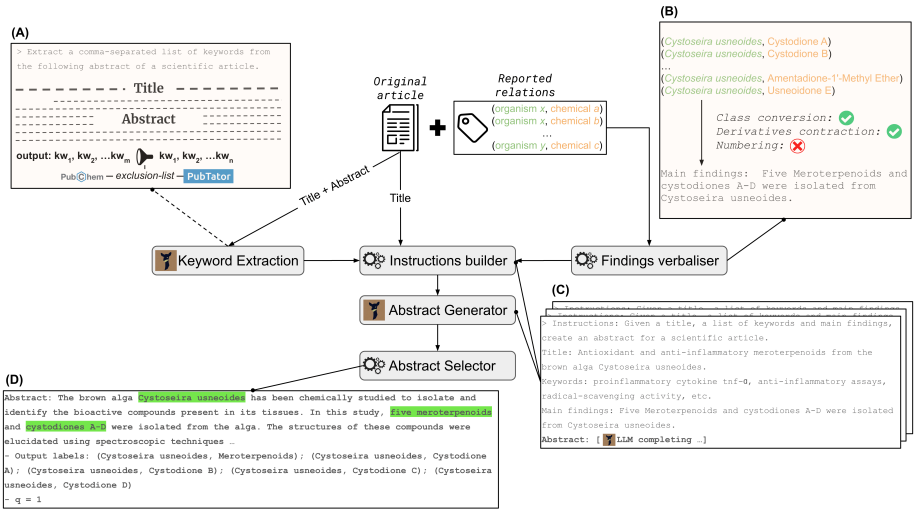


Figure 3
Description of the synthetic abstract generation workflow.

findings. To alleviate this potential issue, an exclusion list is created for each input seed abstract, including organisms and chemicals annotated by LOTUS, their synonyms from PubChem, and annotations from PubTator (Wei et al. 2019). Then, all extracted keywords matching items from this list are excluded.

By explicitly formalizing the expected patterns in upstream instructions, the expression of NP relationships during the generation step can be more efficiently controlled. The findings-verbalizer module operates as a sampler to emulate and combine various patterns of expression that can be observed in the literature. It incorporates 5 possible transformations: (1) members of a same chemical class[3] can be replaced by the simple mention of the class (e.g., a list of chemicals $c_{1:5}$ is replaced by the more concise mention "*Five Meroterpenoids*"); (2) Lists of chemical derivates can be contracted (e.g., "*Cystodione A–D*"); (3) The order of relationships is systematically shuffled; (4) Chemicals can be numbered (e.g., "*Cystodione A-D* (1–4)"); (5) Directions of the relationship can change from "*O produces C*" to "*C was isolated from O*". See Box B of Figure 3 and more details in Appendix A.6. These different transformations are reminiscent of the strategies commonly used in data augmentation (Feng et al. 2021).

For each input seed abstract, m instructions are sampled and assembled following this procedure and forwarded to the model for generation (Box C). See illustrative examples of abstract generation in Appendix A.7. Finally, the selector module selects a top k, from the m generated abstracts, ensuring that at least a proportion q of the expected relations have the labels of the involved organisms and chemicals explicitly mentioned in the generated abstract (Box D). Regarding the expected output labels, the replacement operated by transformation (1) also applies: The initial relations $r_{1:5}$ involving the 5 meroterpenoids are replaced by a single relation r_6 involving "Meroterpenoids" as chemical entity. In contrast, transformation (2) does not affect the output

3 The chemical classes of a compound are determined according to NP-classifier (Kim et al. 2021) annotations in LOTUS.

labels, requiring the model to expand the list of relations involving each derivative (see Box D - *Output labels*). Also, the loss in Equation 3 (like in Seq2rel) is permutation sensitive, but the order created by the transformation (3), which also applied to the output labels, is almost systematically respected by the model in the generated abstract, alleviating this issue. Transformations (4) and (5) have no influence on the output labels.

3. Empirical Experiments

3.1 Imbalanced Repartition of Reported Relations and Coverage on Biological Kingdoms

As reported in the original release of the LOTUS dataset (Rutz et al. 2022), the imbalance in the data distribution manifests at two main levels: the repartition of the number of reported relationships per organism (respectively, chemicals) and the coverage of biological kingdoms. These observations were reproduced from the latest available snapshot of the LOTUS dataset (v10-01-2023),[4] containing more than 533,000 distinct relations between organisms and NP, reported from more than 88,000 articles. As expected, a small fraction of the organisms (respectively, chemicals) attracts a large proportion of the relations: more than 72% of relations involve only 20% of the organisms (Figure 4.A). Beside these Pareto distributions (Newman 2005), the imbalance in the repartition of the relations across biological kingdoms is also important: 80% are related to *Archaeplatida* (Figure 4.B top-left). Considering these two biases is essential to extract a valuable sample. This motivated the use of the GME-sampler in a stratified way, to maximize diversity and reduce the Pareto effect, while ensuring a more balanced coverage across biological kingdoms.

3.2 Dataset Pre-processing

The original dataset was first preprocessed and filtered prior to sampling to eliminate various sources of perturbations and unusable data in subsequent steps. Specifically, only documents with publicly available abstracts on PubMed were selected, and these were further filtered based on the number of reported relations. Indeed, a manual inspection of a subset of articles revealed that documents reporting large numbers of relations (Swainston et al. 2016; Thiele et al. 2013; Stefanini et al. 2017; Thompson et al. 2006) often propose genome-scale metabolic reconstructions, large screening analyses, or database releases. Although these documents may report hundreds of relationships, they are typically not expressed in the abstracts, making them useless examples for building a RE model. Only articles reporting fewer than 20 relations (corresponding to the quantile 93%) were then selected. Compared to organism names, the length of chemical names can exhibit extreme variability and exceed hundreds of characters depending on the nomenclature. To mitigate the issues posed by these lengthy labels, which are inordinate to decode and could consume an excessive portion of the context window during training and testing, only relations involving chemicals with a label length $l \leq 60$ characters were retained. See more details in Appendix C.1 and the global pre-processing statistics in Table C.1. The kingdom coverage is also presented in Figure 4.B top-right.

4 `https://zenodo.org/record/7534071`.

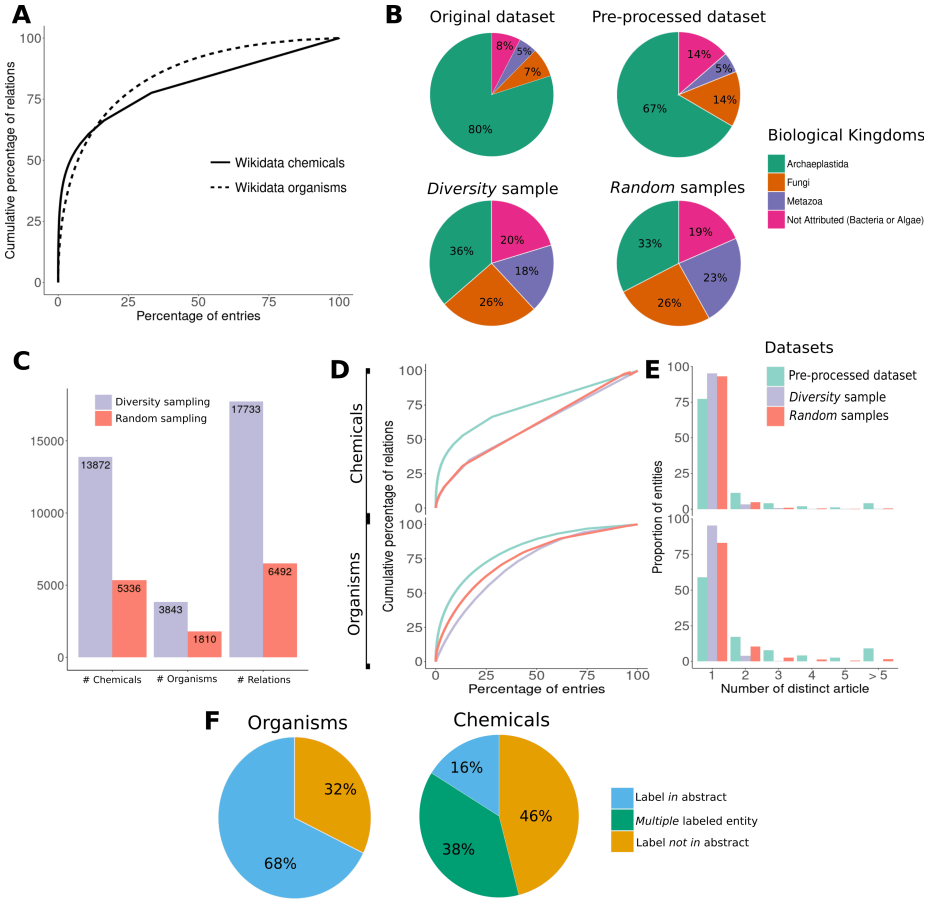


Figure 4
A: Distribution of the cumulative proportion of reported relations per fraction of chemicals (plain line) and organisms (dashed line), ordered by their contributions. The same relations but reported from different articles are considered as distinct items. From both curves, it can be estimated that 20% of the most represented chemicals hold more than 68% of the relations, and 20% of the organisms hold more than 72%. **B:** Repartition of the number of reported relations, organized per kingdom of the subject organism in 4 datasets: the original (*full*) dataset, the pre-processed dataset, the *Diversity* sample, and *Random* samples. For *Random* samples, proportions are averaged over 5 samples. **C:** Statistics of the number of distinct organisms, relations, and chemicals in the *Diversity* sample compared to *Random* samples. All samples contain 2,000 articles. **D:** Similarly to **A**, the distribution of the cumulative proportion of reported relations per fraction of chemicals (top) and organisms (bottom), ordered by their contributions in three different type of samples: *Pre-processed*, *Diversity*, and *Random*. **E:** Distribution of the frequency of mention in distinct articles of chemicals (top) and organisms (bottom). **F:** Mismatches between standardized labels of organisms and chemicals and their original literal mentions in the abstracts of articles reporting the relationships. "Multiple labeled" entities correspond to multiple co-joined chemical mentions that are not expressed in a continuous string. See details in Appendix C.2.

3.3 Building a Diversity-augmented Dataset

3.3.1 Diversity-sampling on Organisms and Chemicals. The preprocessed dataset was first stratified according to the taxonomic classification (kingdoms) of the organisms associated with the relations reported in each document. Subsequently, the GME-sampler was

applied to each subset (Figure 5 Top) to monitor the evolution of the diversity metrics ($H_S(O)$ and $H_S(C)$) and determine an optimal sample size. Indeed, the GME-sampler operates as a ranking method, where the article selected at step n, is the one which contributes the most to the diversity of the set of the $n-1$ articles selected upstream. For both organisms and chemicals, diversity increases rapidly in the first hundred ranked items, followed by a plateau. Specifically for organisms (regardless of the kingdom), diversity showed a decline in the second half of the sampled items (Appendix Table C.2). This is the signal that the addition of new articles provide relations for already well-covered organisms and disrupted the existing balance in the organism distribution. In contrast, the impact of newly added articles on chemicals is negligible, likely because they represent a larger set of distinct entities. To keep a reasonable balance between diversity and sample size, we decided to only retain the top $n = 500$ ranked articles per kingdom, ensuring at least 80% of the maximal observed entropy on both organisms and chemicals (Figure 5 Bottom). The proportions of maximal observed entropy at alternative sample sizes are presented in Appendix Table C.3.

The impact of the diversity-sampling strategy is evaluated by comparing the composition of the sample against 5 random samples of equivalent sizes.[5] The original diversity sample and the extracted random samples are respectively denoted as *Diversity* and *Random* samples. While showing similar kingdoms' coverage because of the common stratification procedure (Figure 4.B Bottom), the diversity sample is, as expected, significantly richer in terms of distinct number of chemicals, organisms, and relations (Figure 4.C). This improved diversity is also reflected in a reduced pareto effect for the distribution of the organisms (negligible for chemicals), and overlap between the entities reported in each article (Figures 4.D and E).

The diversity-sampling strategy was also evaluated against three alternative baselines. In *Top-organisms*, the top 500 articles with the most distinct organisms (individually) were extracted per biological kingdoms. This was similarly done for relations and chemicals with *Top-relations* and *Top-chemicals*. As expected, the *Top-relations* strategy led to the largest set of distinct relations (Figure 5.B), followed by *Top-chemicals* and the proposed diversity-sampling. However, this improvement comes at the expense of a poorer diversity in terms of organisms, but also balance in their distribution (Figures 5.C and D). Interestingly, the *Top-organisms* strategy led to a smaller set of entities compared to the diversity-sampling. Indeed, in the case of an imbalanced distribution of entities over the sampled items (i.e., some model organisms attract more articles than non-model organisms), the simple *Top-organisms* strategy does not consider this potential redundancy. However, its prevention is an explicit objective with the proposed approach. Overall, the evaluated metrics suggest that the diversity-sampling with the GME-sampler offers a valuable compromise between these alternative strategies.

3.3.2 Distance Between Standardized Annotations and Original Text. Several studies emphasize the importance of data quality over quantity for fine-tuning language models (Zhou et al. 2023; Dettmers et al. 2023; Li et al. 2023). LOTUS data are recognized as being of high quality, particularly because of the harmonization, cleaning, and validation steps of the workflow, aligning original records from several open NP databases into standardized structures and organisms in Wikidata. Although this is essential to ensure data FAIRness, these processes logically distance the standardized entries from their original literal mentions in the referenced articles. To get a rough estimate of this

5 Each random sample is composed of 500 random literature items sampled per kingdom.

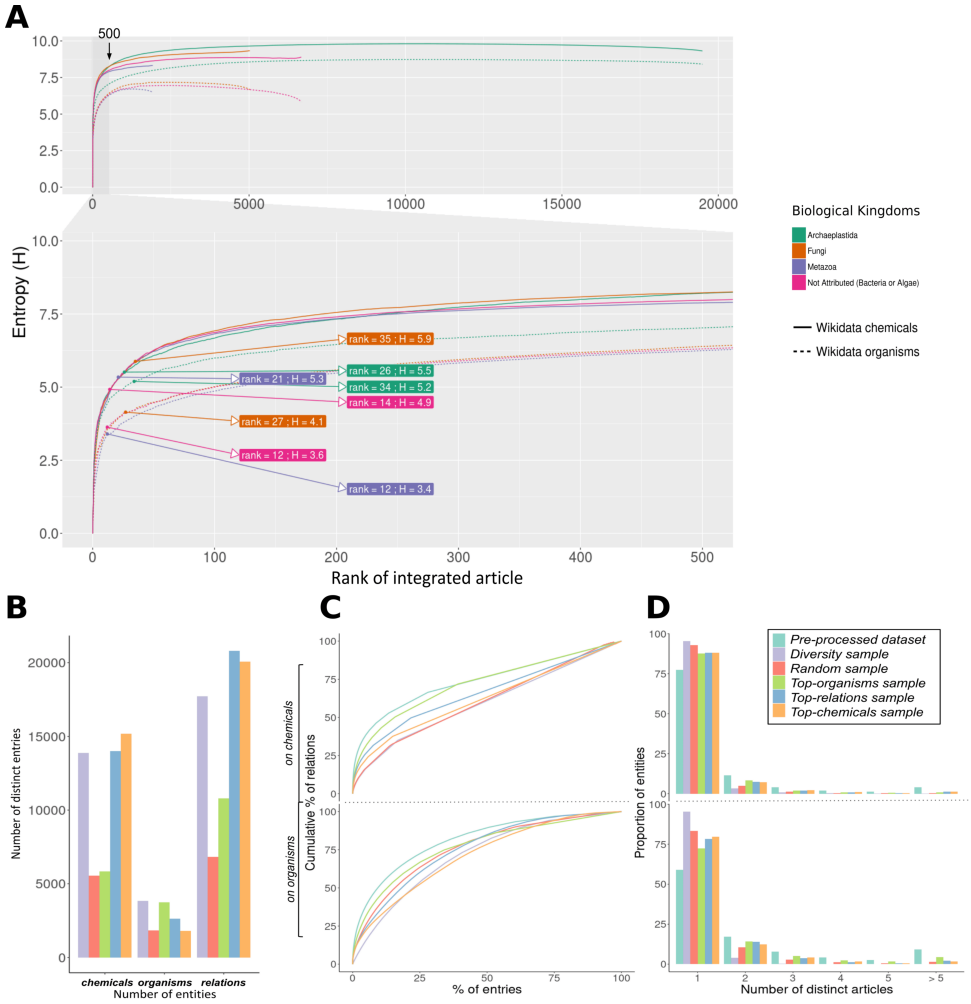


Figure 5
A-*Top-panel:* Evolution of the entropy metric over the distribution of reported organisms ($H_S(O)$) and chemicals ($H_S(C)$) by adding iteratively a new article ($d*$) in the built dataset, stratified by biological kingdom. The step (500) when 80% of the maximal entropy is reached in all the kingdoms, for organisms and chemicals, is indicated with the black arrow (more details in Appendix Table C.3). **A**-*Bottom-panel*: Zoom of the evolution of $H_S(O)$ and $H_S(C)$ in the first 500 added articles. For each curve, the knee-points (*bending* points) with the corresponding ranks and associated entropies are indicated. Below are presented additional evaluation statistics (similar to Figure 4 C, D and E) for the diversity-sampling method compared to other potential sampling objectives. In "*Top-organisms sample*," the top 500 articles with the most distinct organisms (individually) were extracted per biological kingdoms. Similarly, for relations and chemicals, with "*Top-relations sample*" and "*Top-chemicals sample*," respectively. **B:** Statistics of the number of distinct organisms, relations, and chemicals in the *Diversity* sample compared to alternative samples. All samples contain 2,000 articles. **C:** Distribution of the cumulative proportion of reported relations per fraction of chemicals (top) and organisms (bottom), ordered by their contributions in the different samples. **D:** Distribution of the frequency of mention in distinct articles of chemicals (top) and organisms (bottom) in the different samples.

distance, the *Diversity* and the 5 *Random* samples were merged into a single *Extended* dataset. Then, we estimated the proportion of the labels of the standardized entities that could be found in the original abstracts of articles reporting the relationships. Details of this estimation are in Appendix C.2. More than 2/3 of the organism labels are effectively retrieved in the original abstract, while less than half of the chemical names can be retrieved, even considering their synonyms (Figure 4.F). Assuming that these two types of mismatches are independent, only 1/3 of the reported pairs would be completely found in an abstract. Finally, some reported NP relationships are simply not expressed in the abstract of the cited reference, but have been reported from the body of the article or supplementary materials.[6] Whether they are derived from the *Diversity* or a *Random* sample, these noisy examples make the training of a model challenging because some labels to be predicted are missing from the input text (Northcutt, Jiang, and Chuang 2021; Jain et al. 2020). In this context, alternative strategies like zero-shot or few-shot learning (also called in-context learning) based on open LLMs (Liu et al. 2022; Chen et al. 2022b) also need to be considered.

3.3.3 Creating a Manually Curated Evaluation Dataset. If these discrepancies certainly affect the training of a model, they are a more sensitive issue in an evaluation set (Northcutt, Athalye, and Mueller 2021). Also, if diversity can be an important feature for a training set (Yu, Khadivi, and Xu 2022), it is arguably also important for an evaluation set (Liang et al. 2022). While smaller by design, the evaluation set needs to be representative. Finally, because the manual curation of an evaluation set is an expensive and time-consuming task, the selected set of entries need to be chosen carefully (Sambasivan et al. 2021). Considering the last points, the knee-points of the entropy curves (where the entropy increases weaker by new added articles) obtained with the GME-sampler suggest relevant tradeoffs between sample size and diversity (Figure 5 Bottom-panel) early in the sampling. Nonetheless, as they vary between different biological kingdoms, on organisms and chemicals, and could be too restrictive, the extended set of the top 50 items from each kingdom was extracted, resulting in an evaluation dataset of 200 abstracts. The abstracts were manually curated by an expert, annotating all instances of mentioned organism–NP relationships in their order of appearance in the original text and using established identifiers such as Wikidata IDs and PubChem IDs. As isolated chemicals are sometimes grouped into chemical families for the sake of brevity in abstracts, all mentions of a more general chemical family were also annotated. The curated evaluation set is publicly available at `https://zenodo.org/records/8422007`. Details about the curation protocol are available in Appendix B.1, together with a comprehensive overview of the content of the dataset in Appendix B.2. Additionally, we computed the inter-annotator agreement for the organism–NP relationships, based on a separate set of annotations provided by a second annotator using the same guidelines, and achieved 88.5%. Details in Appendix B.3.

3.4 Few-shot Learning Approaches the Performance of Standard Fine-tuning on Raw Data

The mismatches between the standardized labels and the original abstracts have therefore been corrected for the evaluation set. However, due to the considerable investment

6 However, we have chosen to focus only on abstracts and not on full texts because of their much greater availability and their synthetic forms.

Table 1
Performance of 5-shot *in-context* learning using LLaMA and LLaMA derived instructions-tuned models compared to Seq2rel, GPT-2, and BioGPT fine-tuned models. Three types of training datasets are evaluated: the diversity sample (`Diversity-raw`), 5 random samples (`Random-raw`), and the extended sample (`Extended-raw`), which is the union of the previous samples. `Full` is a dataset that contains all available examples from the LOTUS snapshot, except the 200 used in the evaluation set. Best performance via fine-tuning are bold, while best performance in few-shot settings are underlined.

Model	Training	Precision	Recall	F1
LLaMA-7B	Few-shot learning (5-shot)	27.0	9.0	13.6
LLaMA-13B		35.6	<u>23.6</u>	28.5
LLaMA-33B		38.5	23.2	29.0
LLaMA-65B		<u>40.2</u>	23.0	<u>29.2</u>
Alpaca-7B		15.1	2.2	5.9
Vicuna-13B		38.4	20.4	26.5
Seq2rel	`Random-raw`	43.2 +/− (6.7)	4.8 +/− (1.2)	8.6 +/− (2.0)
	`Diversity-raw`	39.6	5.4	9.5
	`Extended-raw`	47.3	5.8	10.4
	`Full`	45.6	7.1	12.2
GPT-2	`Random-raw`	32.5 +/− (4.8)	11.8 +/− (5.3)	15.0 +/− (2.5)
	`Diversity-raw`	22.3	19.2	20.6
	`Extended-raw`	44.8	21.7	29.3
	`Full`	**47.5**	22.5	30.5
BioGPT	`Random-raw`	47.2 +/− (4.0)	19.8 +/− (2.7)	27.6 +/− (2.5)
	`Diversity-raw`	37.1	**28.4**	32.2
	`Extended-raw`	42.2	26.5	**32.5**
	`Full`	46.7	21.3	29.3

of time and resources required for this task, the same corrections were not applied on the remaining data available for training. In this particular context of noisy data for end-to-end RE, two strategies were evaluated: standard fine-tuning and few-shot learning, the latter being able to rely only on a few manually selected examples. The performance of the fine-tuning strategy was evaluated using train/valid datasets derived from the initial *Diversity*, *Random*, and *Extended* samples, which will be referred to as `Diversity-raw`, `Random-raw`, and `Extended-raw`, respectively. Specifically, `Extended-raw` is an extension of `Diversity-raw` that also includes all examples from the 5 `Random-raw` datasets. To further evaluate the impact of dataset size on training performance, models were also trained on the `Full` dataset. The `Full` dataset is larger than `Extended-raw` and contains all available examples from the pre-processed LOTUS dataset (excluding the 200 used in the test set). Their respective sizes and splits are detailed in Appendix Table C.4. All datasets were used to train 3 models for end-to-end RE: Seq2rel, BioGPT, and GPT-2. Six open LLMs were also evaluated in few-shot learning settings: LLaMA 7B, 13B, 33B, and 65B, along with two models, respectively fine-tuned on instructions and conversations and derived from LLaMA 7B and 13B: Alpaca-7B and Vicuna-13B.

Best performance in fine-tuning settings was achieved by BioGPT (Table 1). Regardless of the training dataset,[7] it consistently outperformed Seq2rel and GPT-2 and demonstrated a F1-score of 32.5% when trained on `Extended-raw`. We also evaluated

7 With the exception of the instance trained on the `Full` training set.

the influence of the different training datasets on models performance. The results indicate that models trained on `Diversity-raw` outperformed[8] those models trained on `Random-raw`, with a notable improvement in recall at the expense of precision. Merging the datasets into a larger (`Extended-raw`) also resulted in improved performance for all models. However, expanding the dataset to all available examples only barely improved the previous performance and surprisingly underperformed with BioGPT. In few-shot learning scenarios, the best performance was obtained with LLaMA-65B and declines with smaller models. Although the performance was inferior compared with fine-tuned alternatives, the models achieved reasonable scores considering the limited number of archetypal examples provided. These results also emphasize the potential of few-shot learning or prompt-tuning based approaches in practical context with low-resources.

3.5 Reversing the Task: Generation of Synthetic Data with Open LLMs

While LLMs cannot compete in terms of performance with fine-tuned approaches in the evaluated settings, their generative abilities could be used alternatively to address the main bottleneck: the discrepancies between the input text and the labels in the training data. It requires going beyond distant supervision or data augmentation (Feng et al. 2021; Shang et al. 2018; Smirnova and Cudré-Mauroux 2018). The former involves mapping relationships from a knowledge base to a large corpus of text to generate pseudo-labels, whereas the latter entails applying a range of transformations, permutations, or morphings to a core set of high-quality examples. The semantic discrepancy between the input text and the output labels would not be resolved by introducing syntactic or lexical variations in the original abstracts. Moreover, the results presented in Table 1 indicate that the inclusion of more training (noisy) instances (`Full` dataset) does not result in systematic improvements. In contrast, the adaptive described approach proposes to generate a set of new synthetic input abstracts from a pre-defined context and a set of expected output labels (i.e., organism–NP relationships).

To maintain consistency, each synthetic abstract is based on the context and results reported from an original seed abstract. The first step is to generate the instructions to prompt the selected LLM for generation. The instructions are composed of a title, a list of keywords, and the verbalized main findings (Method 2.3). We decided to use the open source Vicuna-13B (Chiang et al. 2023),[9] a LLaMA-13B model fine-tuned on user-shared conversations collected from ShareGPT,[10] which outperforms alternatives of equivalent sizes on several benchmarks (Dettmers et al. 2023). For each input seed abstract, the top-10 extracted keywords were used in the built instruction. As this is a crucial step, the performance of Vicuna-13B to extract keywords have been evaluated on the SemEval2017-Task10 dataset (Augenstein et al. 2017) in Appendix C.4. To diversify the generated abstracts, $m = 10$ instruction prompts with different verbalization patterns were then sampled per initial seed article. Finally, only the top $k = 3$ most relevant synthesized abstracts per seed were selected with the simple, yet effective, selector module.

To evaluate the impact of diversity-sampling on the seed articles used for synthetic generation, we created two new datasets: `Diversity-synth` and `Random-synth`, derived

8 Measured with F1-score, because it penalizes models with unbalanced performance between recall and precision.

9 version v1.3 from 22/06/2023: `https://huggingface.co/lmsys/vicuna-13b-v1.3`.

10 `https://sharegpt.com/`.

from the original abstracts in the Diversity-raw and Random-raw datasets, respectively. Several illustrative examples of synthetic abstracts from Diversity-synth are discussed in Appendix A.7, highlighting both the variability and the potential caveats (errors, hallucinations) of the process. As with the original data, Diversity-synt and Random-synt were merged in Extended-synt to measure the impact of the dataset size. Statistics of the generated datasets are presented in Appendix Table C.4. In total, more than 25,000 synthetic abstracts were generated from the 7,901 originally contained in the raw datasets. From Diversity-raw, 200 initial items were excluded by the selector module and 162 on average for Random-raw. While the distinct numbers of entities/relations dropped in synthetic datasets, the selector guarantees that these labels are part of the generated abstracts. Furthermore, the generation process enables the integration of examples with chemical classes in the input text and expected labels, which were not available in the original data.

3.6 Training on Synthetic Data Improved Performance over Noisy Raw Data

The synthetic datasets were used to train new instances of the previously evaluated models: Seq2rel, GPT-2, and BioGPT. Although the synthetic training sets (Diversity-synt and Random-synt) are almost half the size of Extended-raw (respectively, 3,562 and 3,798 compared with 7,111 examples), on which was established the previous baseline with BioGPT (F1-score = 32.5), all the trained models demonstrated improved performance (see Table 2). Indeed, all metrics improved in all 9 configurations—3 models × 3 categories of dataset (Random, Diversity, and Extended)—with the best gains observed for Seq2rel. The ranges of improvements for precision, recall, and F1-score go respectively from: 6.2 to 21.9, 13.2 to 25.3, and 12.4 to 30.6. The ranking of the models and the impact of the synthetic training sets on the final performance align with the previous observations on the original data. BioGPT models consistently outperformed Seq2rel and GPT-2, and the training on Diversity-synt resulted in an improved recall at the expense of precision compared with Random-synt. However, the GPT-2 models trained on

Table 2
Performance of Seq2rel, GPT-2, and BioGPT models fine-tuned on synthetic data. Three types of synthetic training datasets are evaluated: the diversity sample (Diversty-synt), 5 random samples (Random-synt), and the extended sample (Extended-synt), which is the union of the previous samples, all synthetically generated from the corresponding seed original (x-raw) samples. For *Random-synt* samples, results are averaged and standard deviations are reported. Best performances are bold, and second best performances are underlined. The absolute improvement on all the metrics (precision, recall, and F1-score) for the models (Seq2rel, GPT-2, and bioGPT) on the 3 types of datasets (Random, Diversity, and Extended) are indicated on the right.

Model	Dataset	Precision	Recall	F1
	Random-synt	62.4 +/− (1.0) (↑ 19.2)	26.8 +/− (2.0) (↑ 22.0)	37.5 +/− (1.9) (↑ 28.9)
Seq2rel	Diversity-synt	61.5 (↑ 21.9)	30.7 (↑ 25.3)	40.1 (↑ 30.6)
	Extended-synt	**65.1** (↑ 17.8)	29.9 (↑ 24.1)	41.0 (↑ 30.6)
	Random-synt	42.6 +/− (2.9) (↑ 10.1)	32.7 +/− (2.8) (↑ 20.9)	37.2 +/− (2.8) (↑ 22.2)
GPT-2	Diversity-synt	28.5 (↑ 6.2)	39.4 (↑ 20.2)	33.0 (↑ 12.4)
	Extended-synt	52.0 (↑ 7.2)	<u>44.6</u> (↑ 22.9)	<u>48.0</u> (↑ 18.7)
	Random-synt	56.4 +/− (2.3) (↑ 9.2)	38.8 +/− (1.9) (↑ 19.0)	46.0 +/− 1.1 (↑ 18.4)
BioGPT	Diversity-synt	52.5 (↑ 16.0)	41.2 (↑ 13.2)	46.2 (↑ 14.4)
	Extended-synt	<u>63.7</u> (↑ 21.5)	**46.5** (↑ 20.0)	**53.8** (↑ 21.3)

Table 3
Performance of fine-tuned BioGPT-Large models on the `Diversity-synt` and `Extended-synt` datasets.

Model	Dataset	Precision	Recall	F1
BioGPT-Large	`Diversity-synt`	57.5	**56.9**	57.2
BioGPT-Large	`Extended-synt`	**69.0**	51.6	**59.0**

`Random-synt` on average outperformed the one trained on `Diversity-synt`, a departure from the trend observed with Seq2rel and BioGPT. Again, the best performance is achieved by BioGPT trained on the merged set, with F1-score = 53.8.

Finally, two BioGPT-Large models were trained on the `Diversity-synt` and `Extended-synt` (see Table 3). The model trained on `Diversity-synt` achieved F1-score = 57.2, comparable to the new best model trained on the much larger merged set (F1-score = 59.0) and also demonstrated a better recall (56.90 against 51.6).

4. Discussion

The application of deep learning models for the completion of biomedical knowledge bases is largely limited by the availability and quality of domain-specific labeled data (Liang et al. 2022). Therefore, we adopted a data-centric methodology (Mazumder et al. 2023; Zha et al. 2023). In order to address the data imbalance and optimize the manual curation process, we proposed the GME-sampler inspired by diversity metrics commonly used in ecology. The sampler was applied on the pre-processed LOTUS dataset (separately on each biological kingdom) to extract a subset of documents, ensuring a diverse set of organisms and chemicals in the reported relations. The compositional analysis revealed a higher number of distinct entities in the extracted sample, but also a better balance considering the fixed number of items. Diversity has been recognized as an important factor in a training set for representation learning and improving the generalization performance of models (Gong, Zhong, and Hu 2019; Yu, Khadivi, and Xu 2022), essential for the NER sub-task. By forcing diversity into the relation partners (organisms and chemicals), we also expect it to be improved in their mentioning contexts. Considering the time and domain expertise requirements to annotate an evaluation dataset, the diversity metric was also used for partitioning. We extracted and manually annotated a representative subset by extracting the 200 top-diverse items. We hope that this manually curated evaluation dataset will help the community to build upon this work.

Despite a smaller number of trained parameters, BioGPT and GPT-2 fine-tuned with QLoRa clearly outperformed Seq2rel. This highlighted the benefit of the larger pre-training, but also the effectiveness of the QLoRA strategy, where low-rank updates of a large, but quantized, model achieve better performance than the full fine-tuning of a smaller model, for a lower parameter budget (Aghajanyan, Gupta, and Zettlemoyer 2021; Dettmers et al. 2022; Hu et al. 2022). While based on the same architecture, improvements of BioGPT over GPT-2 can be attributed both to the pre-training on PubMed and also to the dedicated tokenizer (see Appendix Figure C.2). Beyond the architectures of the models, the training dataset also had a significant impact on the performance. A comparison between models trained on the largest (`Extended-raw`) and the diversity-optimized dataset revealed that the latter achieved competitive results

despite its smaller size. Additionally, results also suggest that improving the diversity of the provided set of examples for training can improve the recall of the model, at the expense of precision. Intuitively, we speculate that the extensive variety of distinct named entities present in the `Diversity` samples (see Figure 4.C) may benefit the NER sub-task learned by the models. However, an increase in the number of identified named entities and the complexity of the examples (with more entities comes more potential relations) could not be as beneficial for the learning of the second sub-task: RE. This could result in more sensitive models: higher recall but lower precision. Overall performance (measured by F1-score) is improved with the `Diversity` dataset on raw data and is equivalent or better for Seq2rel and BioGPT on synthetic data. Also, increasing the number of training examples with noisy data also have limited benefits, as suggested by the comparison with the (`Full`) training dataset extended to all available data (no sampling, no stratification) (Salhofer, Liu, and Kern 2022; Prusa, Khoshgoftaar, and Seliya 2015; Liang et al. 2022). Finally, few-shot learning techniques leveraging open LLMs exhibit reasonable performance (see LLaMA-65B) and can be particularly valuable when only limited or noisy data are available. However, their larger size may incur higher management costs, necessitating careful consideration of resource allocation.

Instead of using them to directly perform the task, we then propose to use them to generate synthetic examples and alleviate the noise of the dataset. However, evaluating the quality of the generated abstracts is challenging. Although the process is prone to hallucinations, the factuality is not the key criteria, as long as the generated texts are credible, meaning that they are coherent and adhere to the established syntax, style, and patterns of expression of the relations in human-written abstracts. Since the training sets of LLMs contain scientific articles and abstracts, they have absorbed their stylistic and syntactic specificities. The generation of synthetic data could then be seen as a form of knowledge distillation. Moreover, while previous studies have suggested that LLMs may not be knowledgeable (Cao et al. 2021; Si et al. 2023; Mallen et al. 2023), other investigations have highlighted the remarkable capabilities of chatbot and instructions-tuned models in following style instructions (Pu and Demberg 2023; Chia et al. 2024). Then, a first relevant evaluation criteria for these synthetic data is the improvement on the performance they provided. Additionally, we measured the textual similarities between synthetic data and original abstracts from the natural products' literature with an n-gram overlap analysis in Appendix A.8. The impact of hallucinations (more precisely instruction inconsistencies) on synthetic data and performance of trained models is also evaluated.

All three models, with different architectures or pre-training data, demonstrated improvements across all metrics, on the 3 categories of datasets (`Random`, `Diversity`, `Extended`), highlighting the benefits of synthetic data in contexts of initial sparse labeled data. Also, transitioning from raw noisy data to synthetic data did not alter the previously observed trend: BioGPT outperformed other models, and the diversity-optimized sampling had a positive effect on the recall of trained models when used to select the seed articles. Most importantly, we noticed that the transition from original to synthetic data had a more determinant impact on the performance improvements than the choice of the model architecture (Seq2rel, GPT-2, BioGPT). For instance, the influence of synthetic data on the performance of BioGPT and GPT-2 is greater than the difference between the two fine-tuned models. The performance of Seq2rel was also enhanced almost by a factor of 4, notably narrowing the gap with GPT models. Similarly, scaling-up the architecture with Bio-Large ($> 4.5\times$ larger) indeed resulted in improved performance, but comparable to the previous enhancement obtained with

synthetic data. Also, we noticed a clear impact of the training dataset, with the best observed recall achieved with `Diversity-synt`. These results also support the data-centric view, even in low-resource scenarios, by demonstrating improved performance over other strategies such as few-shot learning (Xu et al. 2022).

5. Limits and Future Work

Fine-tuned methods exhibit superior performance compared with zero-shot/few-shot approaches. However, the basic prompting approach used in the experiments may not fully demonstrate the capabilities of the models, and alternative strategies have been proposed (Zhao et al. 2021; Wu et al. 2023; Liu et al. 2022). Nevertheless, Jimenez Gutierrez et al. (2022) noted that even with these improvements, the models still lack the accuracy of fine-tuned approaches with qualitative data. The use of LLMs to generate abstracts also has some evident limitations. The generated abstracts exhibit a narrow range of styles to express the relationships between organisms and chemicals compared to human-written abstracts. Although we argued that strict data augmentation could not effectively bridge the initial gap between text and labels, token or sentence level augmentations on generated abstracts could, however, improve both the quantity and diversity of synthetic data (Chen et al. 2023). We suppose that the synthetic data mostly improved the recognition of organism and chemical entities, this sub-task being inherently embedded in the ultimate task of decoding the relationships. Following Kim et al. (2022), LLMs could also be used to generate alternative demonstrations for *in-context* learning. Nonetheless, such approaches need to be further evaluated in the specific context of the biomedical literature.

Although the proposed framework is effective, it cannot guarantee the true diversity of the generated abstracts and the final selection may be very similar. Secondly, the selector module does not ensure that the relations are semantically expressed in the generated abstracts, as it only checks for the explicit mention of the entities. Finally, all generated examples are designed as "positive" cases, meaning that a relation is always expected, which may not be the case in practical applications. The developed models are intended for use on a large corpus of articles and the input documents can be either selected by an upstream retriever component, or, the predictions can be re-evaluated by a downstream selector. Continuing with this data-centric view, future works will prioritize improving the three key components (instructions builder, generator, and selector) to improve the diversity of the synthetic abstracts, rather than focusing on the architecture of the trained models.

Given the highly dynamic nature of the LLM research area, we anticipate significant advancements in model architecture and accessibility to arise from the research community. At the date of writing, the release of LLaMA (and LLaMA2) has paved the way for the creation of more open-license models, such as the next-generation of Vicuna,[11] Mixtral (Jiang et al. 2024), or PMC-LLaMA (Wu et al. 2024), and BioMistral (Labrak et al. 2024) trained on the biomedical literature. The development of multilingual open LLMs (Scao et al. 2022) also offers opportunities for synthetic data generation in promising areas, such as the extraction of plant-disease relationships from Traditional Chinese Medicine prescriptions, where the scarcity of labeled data is limiting (Li et al. 2022).

11 `https://huggingface.co/lmsys/vicuna-13b-v1.5`.

6. Conclusion

With the aim of assisting the completion of NP databases, we provide the first training and evaluation datasets along with the first trained models for end-to-end RE of relationships between organisms and chemicals. Along with these main results, we explored different strategies and proposed new developments to address the problematics raised in this biomedical context. We empirically showed the benefit of the proposed GME-sampler for building a diverse and balanced evaluation dataset as well as its positive impact on the recall via the training data. The results also indicate that the opportunities brought by the open LLMs in scenarios with little or weakly labeled data may not lie only in their zero/few-shot learning abilities, but also in their great potential as synthetic data generator. They could open the door for the extraction of previously unexplored relationships between biomedical entities expressed in the literature, a prerequisite to unlock new paths of inferences in knowledge discovery.

Appendix A. Experimental Setup and Implementation Details

A.1 Few-shot *In-context* Learning Details

The prompt used for few-shot *in-context* learning with $K = 5$ archetypal input-completion examples with LLaMA (7B, 13B, 33B, 65B) is provided in Figure A.1. We used greedy decoding, setting the temperature to 0. Considering their particular fine-tuning, small adjustments were provided to the prompt for Alpaca-7B and Vicuna-13B. All models were also quantized for memory efficient inferences, and average inference times are presented in Table A.1. Considering our available resources, we were not able to use the `q8` (8 bits) quantization for LLaMA models $> 13B$ and improvements in performance could then be expected. In parallel, we noticed significant performance degradations when using `q4` (4 bits). We used `llama.cpp`[12] for quantization and inferences.

The task is to extract relations between organisms and chemicals from the input text.
INPUT: The antimicrobially active EtOH extracts of Maytenus heterophylla yielded a new dihydroagarofuran alkaloid, beta-acetoxy-9alpha-benzoyloxy-2beta,6alpha-dinicotinoyloxy-beta-dihydroagarofuran, together with the known compounds beta-amyrin, maytenfolic acid, 3alpha-hydroxy-2-oxofriedelane-20alpha-carboxylic acid, lup-20(29)-ene-1beta,3beta-diol, (-)-4'-methylepigallocatechin, and (-)-epicatechin.
OUTPUT: Maytenus heterophylla produces 1beta-acetoxy-9alpha-benzoyloxy-2beta,6alpha-dinicotinoyloxy-beta-dihydroagarofuran. Maytenus heterophylla produces beta-amyrin. Maytenus heterophylla produces maytenfolic acid. Maytenus heterophylla produces 3alpha-hydroxy-2-oxofriedelane-20alpha-carboxylic acid. Maytenus heterophylla produces lup-20(29)-ene-1beta,3beta-diol. Maytenus heterophylla produces (-)-4'-methylepigallocatechin. Maytenus heterophylla produces (-)-epicatechin.
INPUT: Ten new ergosteroids, gloeophyllins A-J (1-10), have been isolated from the solid cultures of Gloeophyllum abietinum.
OUTPUT: Gloeophyllum abietinum produces gloeophyllin A. Gloeophyllum abietinum produces gloeophyllin B. Gloeophyllum abietinum produces gloeophyllin C. Gloeophyllum abietinum produces gloeophyllin D. Gloeophyllum abietinum produces gloeophyllin E. Gloeophyllum abietinum produces gloeophyllin F. Gloeophyllum abietinum produces gloeophyllin G. Gloeophyllum abietinum produces gloeophyllin H. Gloeophyllum abietinum produces gloeophyllin I. Gloeophyllum abietinum produces gloeophyllin J.
INPUT: The present work describes the isolation of the cyclic peptides geodiamolides A, B, H and I (1-4) from G. corticostylifera and their anti-proliferative effects against sea urchin eggs and human breast cancer cell lineages.
OUTPUT: G. corticostylifera produces geodiamolide A. G. corticostylifera produces geodiamolide B. G. corticostylifera produces geodiamolide H. G. corticostylifera produces geodiamolide I.
INPUT: Four new cyclic peptides, patellamide G (2) and ulithiacyclamides E-G (3-5), along with the known patellamides A-C (6-8) and ulithiacyclamide B (9), were isolated from the ascidian Lissoclinum patella collected in Pohnpei, Federated States of Micronesia.
OUTPUT: Lissoclinum patella produces patellamide G. Lissoclinum patella produces ulithiacyclamide E. Lissoclinum patella produces ulithiacyclamide F. Lissoclinum patella produces ulithiacyclamide G. Lissoclinum patella produces patellamide A. Lissoclinum patella produce patellamide B. Lissoclinum patella produces patellamide C. Lissoclinum patella produces ulithiacyclamide B.
INPUT: Chemical investigation of Trogopterus faeces has led to the isolation of seven flavonoids. Their structures were elucidated by chemical and spectral analyses. In an anticoagulative assay, three kaempferol coumaroyl rhamnosides had significant antithrombin activity. This is the first report on the occurrence of flavonoid glycosides in Trogopterus faeces.
OUTPUT: Trogopterus faeces produces flavonoids. Trogopterus faeces produces kaempferol coumaroyl rhamnosides. Trogopterus faeces produces flavonoid glycosides.
INPUT: **[ABSTRACT]**
OUTPUT:

Figure A.1
The prompt used for few-shot learning on LLaMA (7B, 13B, 33B, and 65B) models and containing five archetypical examples.

Table A.1
Table of the applied quantizations on the LLMs. See `https://github.com/ggerganov/llama.cpp`. The inference time (in ms) was also measured in 5-shot *in-context* learning settings on the provided evaluation dataset.

Model	Quantization type (size in GB)	Average inference time in ms (± sd)
LLaMA-7B	q8 (6.8 GB)	38,672 (± 22,418)
LLaMA-13B	q8 (13.2 GB)	74,102 (± 6,955)
LLaMA-33B	q5_K_M (21.9 GB)	143,418 (± 72,740)
LLaMA-65B	q5_K_M (44.1 GB)	238,103 (± 122,970)
Alpaca-7B	q8 (6.8 GB)	32,293 (± 16,099)
Vicuna-13B	q8 (13.2 GB)	67,504 (± 32,642)

12 llama.cpp github repo: `https://github.com/ggerganov/llama.cpp`.

Table A.2
Statistics of the number of trainable parameters per evaluated models.

	Total parameters	Trainable parameters
Seq2rel	118546185	118546185 (100%)
BioGPT	350649472	3886208 (1.11%)
BioGPT-Large	1582722536	11533736 (0.73%)
GPT-2 *Medium*	358381208	3555992 (0.99%)

A.2 Choice of the Models

We selected 3 models for evaluation: Seq2Rel,[13] BioGPT[14] (and its variant BioGPT-Large[15]), and GPT-2.[16] Seq2rel was originally designed for end-to-end RE, and was later outperformed by BioGPT. With this minimal set of models, we aim to evaluate the performance of two distinct architectures: Seq2Rel (encoder-decoder) and BioGPT or GPT-2 (encoder-only). Note that BioGPT and GPT-2 share the same architecture. Additionally, we evaluate two pre-training settings: BioGPT on Pubmed articles[17] and GPT-2 on a non-biomedical corpus. Furthermore, we explore two training approaches: full fine-tuning on Seq2rel and tuning via adapters with QLoRA on BioGPT and GPT-2. The number of trained parameters for each model is detailed in Table A.2.

A.3 Fine-tuning Details

Dettmers et al. (2023) demonstrated the efficacy of the QLoRA approach by showing that the loss in performance due to quantization can be fully recovered through subsequent fine-tuning of the adapters, and that increasing the number of adapters is crucial to match full fine-tuning performance. By exploiting the memory benefits of the `NF4` data type, we applied LoRA adapters to all linear blocks (except the initial embeddings layer) of the BioGPT and GPT-2 models. Details on the number of trained parameters are presented in Table A.2. During training, the special tokens `<BOS>` and `<EOS>` are used to delimitate the input X and the expected linearized output Y, such as [X, `<EOS>` `<BOS>`, Y, `<EOS>`]. The `<BOS>` token triggers the RE task at inference time.

For all evaluated datasets, models were then trained during 15 epochs (10 for BioGPT-Large) with 100 warm-up steps and the best epoch was selected using the validation set. We set learning-rate $= 1e-4$, LoRA-$r = 8$, LoRA-$\alpha = 16$, batch size $= 16$. We used the available implementation of QLoRA with PEFT (Mangrulkar et al. 2022). We used the recommended $8-$ *bits* paged AdamW optimizer[18] (Dettmers et al. 2022).

For Seq2rel, we applied a standard full fine-tunings as in the original article. All the fine-tuning experiments were conducted on an NVIDIA GeForce RTX 3090. See details on hyperparameter tuning in Section A.4.

13 Link to Seq2rel GitHub: `https://github.com/JohnGiorgi/seq2rel`.
14 Link to BioGPT model card: `https://huggingface.co/microsoft/biogpt`.
15 Link to BioGPT-Large model card: `https://huggingface.co/microsoft/BioGPT-Large`.
16 Link to GPT-2 model card: `https://huggingface.co/openai-community/gpt2-medium`.
17 Also, the encoder used for Seq2rel is also PubMedBERT: `https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext`.
18 `https://github.com/TimDettmers/bitsandbytes`.

Table A.3
Hyperparameter values used for BioGPT. Hyperparameters were fine-tuned using Optuna on the *Diversity-synt* dataset. Values between parentheses correspond to adaptation for BioGPT-Large. The following settings were evaluated: `batch size` $\in \{4, 8, 16\}$; `learning-rate` $\in [1e-6; 1e-3]$; `LoRA configurations` $\in \{(r = 4, \alpha = 4), (r = 4, \alpha = 8), (r = 4, \alpha = 16), (r = 8, \alpha = 8), (r = 8, \alpha = 16), (r = 8, \alpha = 32), (r = 16, \alpha = 16), (r = 16, \alpha = 32), (r =, 16\alpha = 64)\}$. `Gradient accumulation steps` values were directly scaled in inverse proportion to the `batch size`: $\{20, 10, 5\}$. For the decoding strategies, the following settings were also evaluated: `beam-size` $\in \{3, 5\}$; `stopping criteria` $\in \{$True, False, never$\}$; `length penality` $\in [0, 3]$.

	Tuned ?	Value
Training		
`Batch size`	yes	16 (12)
`Number of epochs`	no	15
`LoRa r`	yes	8
`LoRa alpha`	yes	16
`Learning rate`	yes	1.00e-4
`Weight decay`	no	0.01
`Gradient accumulation steps`	no*	5
`LoRA dropout`	no	0.05
`LoRa target modules`	no	q_proj, k_proj, v_proj, out_proj, fc1, fc2, output_projection
Decoding		
`strategy`	yes	beam search
`beam size`	yes	3
`stopping criteria`	yes	never
`length penality`	yes	1.5
`temperature`	no	0

A.4 Hyperparameter Tuning

Hyperparameter settings, including learning-rate, batch size, and LoRA config, were evaluated on the `Diversity-synt` dataset with Optuna (Akiba et al. 2019). A summary of the hyperparameters tuned for BioGPT is presented in Table A.3. The F1-score on the validation set was used as evaluation criteria. In line with Giorgi, Bader, and Wang (2022), a greedy decoding approach was utilized during the hyperparameter tuning phase, followed by a fine-tuning of the decoding strategy on the configuration that yielded the best results. The experimental setup involved $n = 140$ trials, each consisting of 5 epochs, and was executed using the TPE (Tree-structured Parzen Estimator) sampler and a median pruner. The results of the hyperparameter optimization are presented in Table A.3.

The relationships between hyperparameters and performance are depicted in panel A of Figure A.2. While the batch size and the LoRA configuration don't show strong impact on the final performance, the learning rate was identified as a critical parameter. With the TPE sampler, the learning-rate of the trials rapidly converged around $1e-4$, resulting in stable performance across different batch sizes and LoRA configurations. The impact of the LoRA rank r is more precisely illustrated in panels B and C. As previously observed by Aghajanyan, Gupta, and Zettlemoyer (2021) and Hu et al. (2022), increasing the rank of the LoRA adapters from $r = 8$ to $r = 16$ resulted in only marginal improvements, considering the doubling of the number of trained parameters. The boxplots in panel C also highlight close performance with small variability on validation F1-score for trials with LoRA $r = 8$ and $r = 16$. After choosing the final

training configuration ($lr = 1e-4, r = 8, \alpha = 16$, batch size $= 16$), the decoding strategy was fine-tuned with 40 trials, evaluating greedy decoding and beam search with beam sizes of 3 or 5 (see panel D). Ultimately, beam search with beam size $= 3$ was selected. The best hyperparameter settings obtained for BioGPT were reused for GPT-2 as they share the same architecture, and later for BioGPT-Large.

Similarly to the panel A in Figure A.2, the same hyperparameter tuning experiments were conducted for Seq2rel (see Figure A.3). It consisted of 30 trials on 10 epochs on the `Diversity-synt` dataset. A summary of the hyperparameters tuned for Seq2rel is presented in Table A.4.

A.5 Evaluation Details

All evaluated models (in fine-tuning and few-shot settings) were evaluated for end-to-end RE, jointly performing NER and RE, framed as a generative task. The performance of the tested models were assessed by measuring the `F1-score` over the predicted

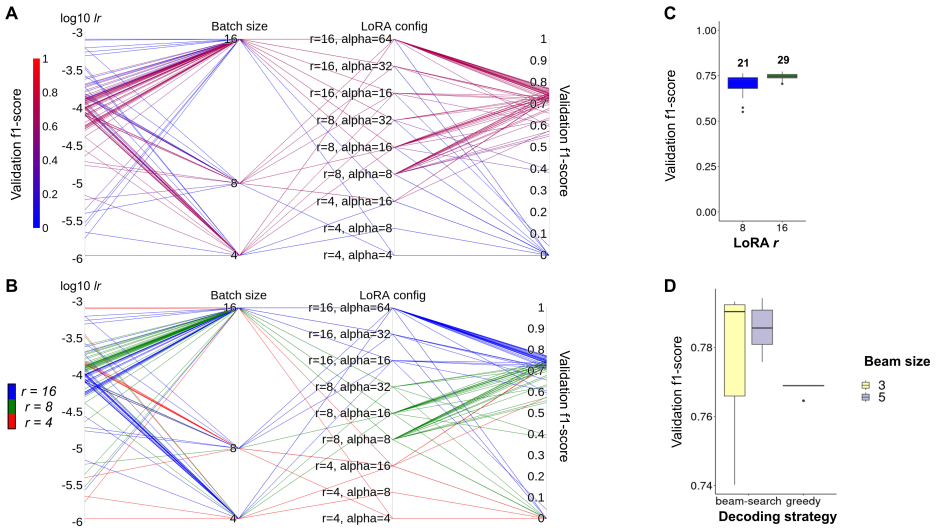


Figure A.2
A: Analysis of hyperparameters for the fine-tuning of BioGPT. The hyperparameter tuning initially consist of 140 trials with the TPE sampler and a median pruner. Only the 96 'completed' trials are considered (44 trials were pruned). The trials were computed with 5 epochs, on the `Diversity-synt` dataset (train and valid). Each line represents a trial, and the color is scaled by the final F1-score obtained on the validation dataset for this trial. While the batch size and the LoRA configuration don't show substantial impact on the final performance, the learning rate (lr) was identified as a key parameter. **B:** Same as **A**, but the lines (representing trials) are colored by the LoRA r parameter (4, 8, or 16). **C:** Boxplot representing the median and variability of the F1-score on the validation datasets from trials using LoRA $r = 8$ or $r = 16$. The number of concerned trials is indicated on the top. **D:** Hyperparameter tuning for the decoding strategy. The boxplots illustrate the distribution of the obtained F1-score on 40 trials for the two evaluated strategies: greedy decoding and beam search, with beam size of 3 or 5. The BioGPT model correspond to the selected configuration ($lr = 1e-4, r = 8, \alpha = 16$, batch size $= 16$). The length penalty and the stopping criteria were also evaluated but did not show significant impact, so they were not included in the figure.

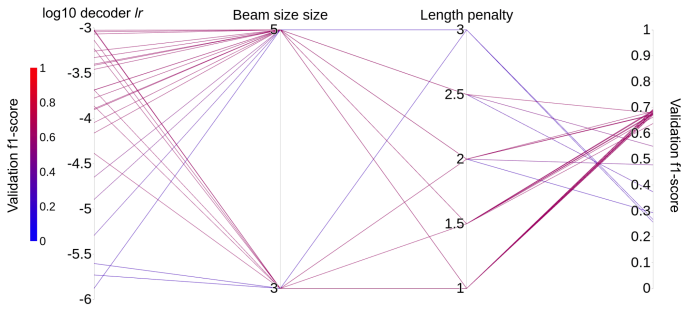


Figure A.3
Analysis of hyperparameters for the fine-tuning of Seq2rel. The hyperparameter tuning initially consist of 30 trials with the TPE sampler. The trials were computed with 10 epochs, on the `Diversity-synt` dataset (train and valid). Each line represents a trial and the color is scaled by the final F1-score obtained on the validation dataset for this trial.

Table A.4
Hyperparameter values used for Seq2rel. Hyperparameters were fine-tuned using Optuna on the *Diversity-synt* dataset. All non-mentioned parameters were set according to the CDR configuration in the original Seq2rel's article. The following configurations were evaluated: decoder's `learning-rate` $\in [1e-6; 1e-3]$; `beam-size` $\in [3, 5]$; `length penality` $\in [1, 3]$.

	Tuned ?	Value
Training		//
decodr learning rate	yes	9.00e-4
batch size	no	4
number of epochs	yes	20
gradient accumulation steps	no	10
others	no	identifical to seq2rel's CDR config
Decoding		//
beam size	yes	5
length penality	yes	1

relations extracted from the decoded outputs. An extracted relation is considered correct only if the head (an organism) and the tail (a chemical) entities exactly match the ground-truth labels.

A.6 Main Findings: Verbalization Patterns

To emulate different patterns of expression of the NP relationships, 5 transformations are applied: (1) chemical class replacement, (2) derivates contraction, (3) shuffling, (4) numbering, and (5) relation directionality. The findings-verbalizer module operates as a sampler, and each transformation has an assigned probability. In the conducted experiments, we used $p_1 = 0.2$, $p_2 = 0.9$, $p_3 = 1$ (systematic shuffle), $p_4 = 0.25$, and $p_5 = 0.9$ for the corresponding transformations. The values were empirically estimated from observed behaviors in the literature. To enhance the diversity of the generated abstracts, the temperature parameter is also randomly sampled in the next generation step: $t \in \{0.5, 0.6, 0.7, 0.8\}$, as similarly evaluated by Chung, Kamar, and Amershi (2023). All other decoding parameters were set by default: top-K = 40, top-P = 0.95, and

repeat-penalty = 1.1. Similarly to few-shot learning, we also used `llama.cpp` through the Python bindings library `llama-cpp-python`[19] for inference in generating the synthetic abstracts. We monitored the generation time and observed that on average[20] a synthetic abstract is produced in 35,708 ($\pm$13,945) ms, showing a significant variability depending on the prompt (min $\approx$ 10s and max $\approx$ 2 min). All generation experiments were conducted on a NVIDIA GeForce RTX 3090.

A.7 Examples of LLM Prompting for Synthetic Abstract Generation

The following section provides archetypal examples to illustrate the diversity engendered by the synthetic abstract generation process. Recall that each generation is calibrated with an original title, a set of keyphrases derived from the original abstract, and verbalized main findings. In the latter, 5 main transformations can be applied to improve the diversity of the generation (see Method 2.3).

These transformations allow for the generation of multiple alternative synthetic abstracts, which emulate different syntaxes or styles for communicating the isolation of the same set of compounds (see Figure A.4). A serves as a reference for a *standard* instruction/generation. The example B introduces variations by reshuffling the order of the mentioned chemicals and then numbering them. In C, different subsets of compounds were substituted with their associated chemical families. In A and B, the expected output labels align with the verbalized main findings, e.g: "*Lachnum papyraceum* produces 6-Methoxymellein; *Lachnum papyraceum* produces 4-Chloro-6-methoxymellein, ...". In C, they are substituted by the chemical classes: "*Lachnum papyraceum* produces Coumarins; etc.."

However, for multiple co-joined chemicals (Figure A.5), while the synthetic text mention "cytosporones J-N, pestalasins A-E", the outputs are expected to be expanded like: "Cytosporone J, Cytosporone K, Cytosporone L, ..., Pestalasin A, Pestalasin B, ..., Pestalasin E". Verbalized relations can also exhibit a N:M pattern, when multiple compounds are isolated from multiple organisms, showcasing the model creative generation abilities (Figure A.6).

The generation process is subject to certain limitations and can occasionally produce inaccuracies of a similar nature to those that were intended to be mitigated. In Figure A.7, while it is explicitly indicated in the instruction part that "*Tagetes erecta* produces two Flavonoids", this information does not appear in the generated abstract. Additionally, the NPs isolated from *Tagetes lucida* are qualified as Flavonoids, which is a wrong assertion, i.e., a hallucination. The synthetic abstracts frequently exhibit instances of hallucinations, yet, these do not significantly impair their utility for the specific task of RE, as long as they do not pertain to the expression of the relationships (see Figure A.8).

A.8 Synthetic Abstracts: Empirical Analysis of *N*-gram Overlap and Impact of Hallucinations

N-gram-based metrics (e.g., BLEU-score) have been widely used to assess the quality of text generation in machine translation and for author-style classification (Papineni et al. 2001; Sidorov et al. 2014; Ríos-Toledo et al. 2022). Intuitively, *n*-grams capture the frequency of words, as well as lexical and syntactic properties of a text. We computed

19 `https://github.com/abetlen/llama-cpp-python`.
20 On the 15190 generations used for the `Diversity-synt` dataset.

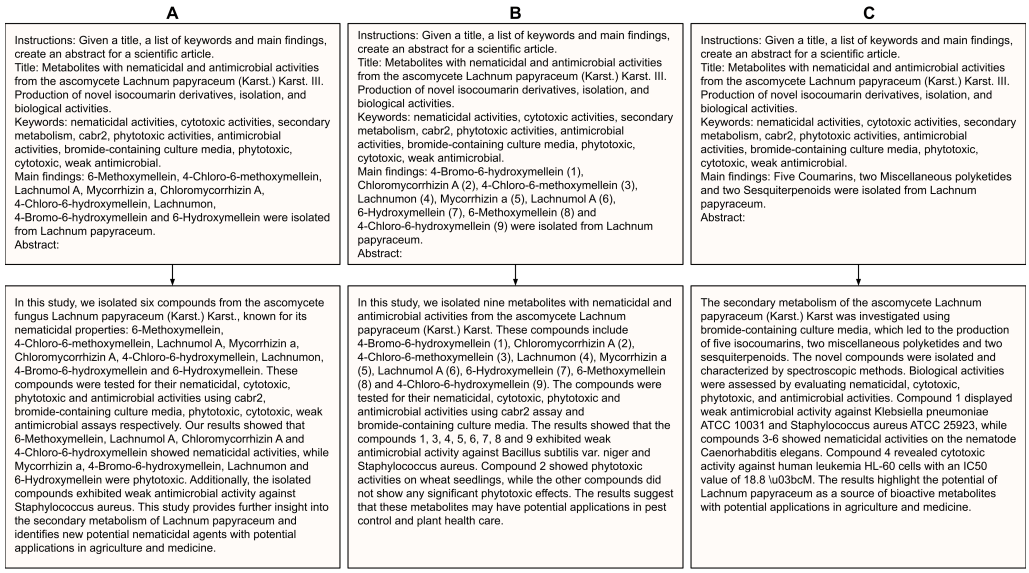


Figure A.4
Three generated abstracts from the seed article PMID: 7730162. A is a standard generation. In B, the chemicals are shuffled and numbered. In C chemicals were substituted with their corresponding chemical families. All generation was produced with $temp = 0.7$ and are training examples from the `Diversity-synt` dataset.

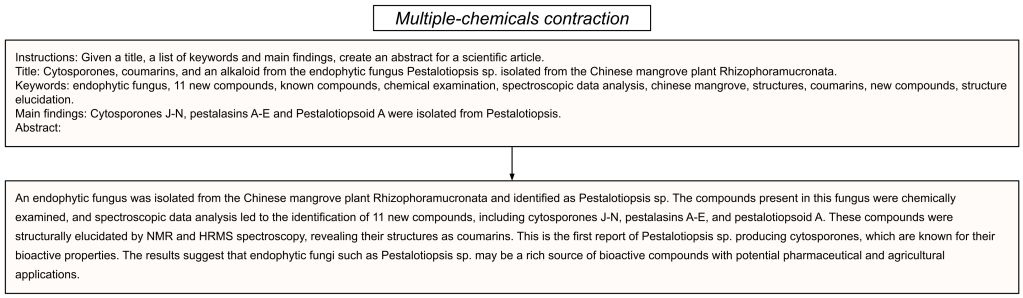


Figure A.5
An example of an abstract generation with multiple co-joined chemicals in the instruction. The original seed article is PMID 19762244. The generation was produced with $temp = 0.7$. It is a training example from the `Diversity-synt` dataset.

the proportion of overlap between the top-50, top-100, and top-500 most frequent word n-grams in the generated abstracts compared to three distinct reference sets: the original seed articles used for generation (*vs. Originals*), random articles sampled from LOTUS (*vs. LOTUS*), and random articles from PubMed (*vs. Randoms*). By comparing the n-gram overlaps, we aim to determine if the generated abstracts are more similar to those from the natural products literature (*Originals* and *LOTUS*) than to random biomedical abstracts from PubMed (*Randoms*). Panel A in Figure A.9 shows a similar proportion of n-gram overlap between synthetic abstracts and the *Originals* and *LOTUS* sets, also consistently better than with random articles. Such frequent and shared n-gram include

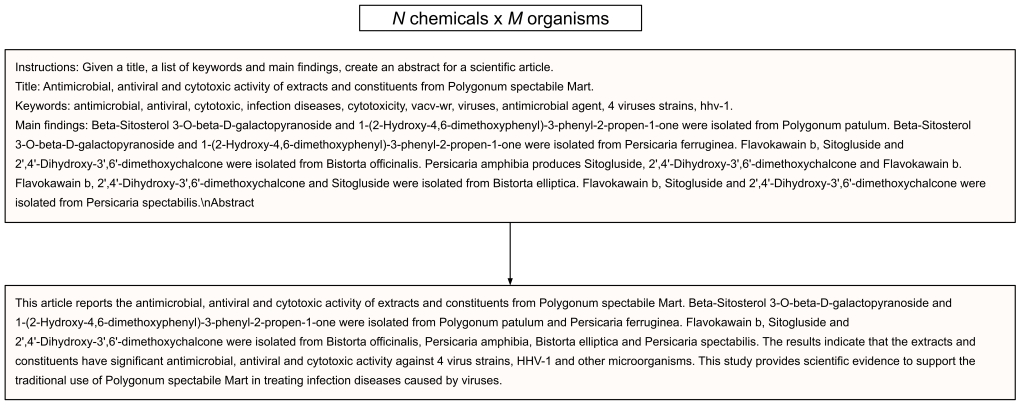


Figure A.6
An example of an abstract generation with an `N:M`-type relations. The original seed article is PMID 20382006. The generation was produced with $temp = 0.7$. It is a training example from the `Diversity-synt` dataset.

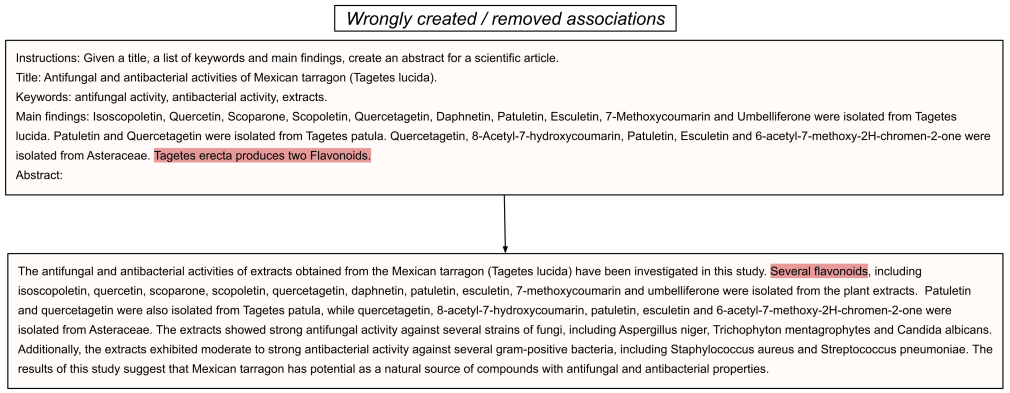


Figure A.7
An example of an abstract generation with incorrectly added and removed relations. The original seed article is PMID 19127719. The generation was produced with $temp = 0.5$. It is a training example from the `Diversity-synt` dataset.

for instance: "*were isolated from*", "*structures were elucidated*", "*1D and 2D NMR*", "*with IC50 values*", etc.

The impact of hallucinations on the quality of synthetic data is also important to consider. We suggested that factual hallucinations on contextual elements in the generated abstracts (e.g., Figure A.8) are less harmful for the quality of synthetic data, than hallucinations related to the expression of the relations in the main findings (e.g., Figure A.7). These hallucinations are classified as *Instruction inconsistency*, when the output of the LLM deviates from the user instructions (Huang et al. 2023).

To evaluate their impact on the performance of trained models, we constructed a new synthetic dataset, based on the same seed abstracts as `Disversity-synt`, but using

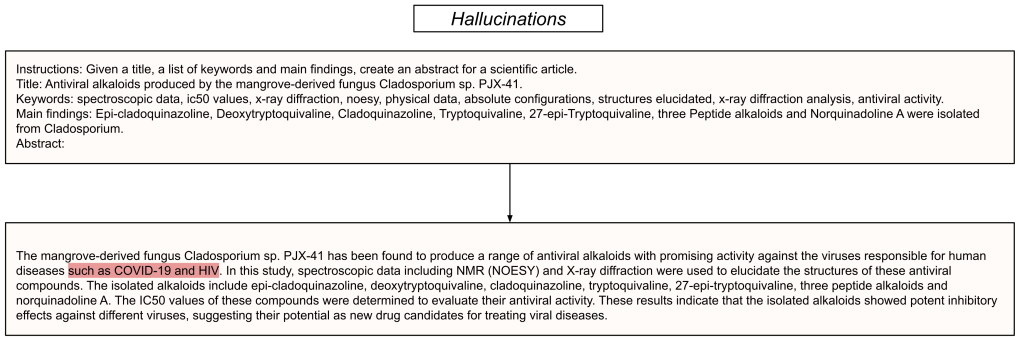


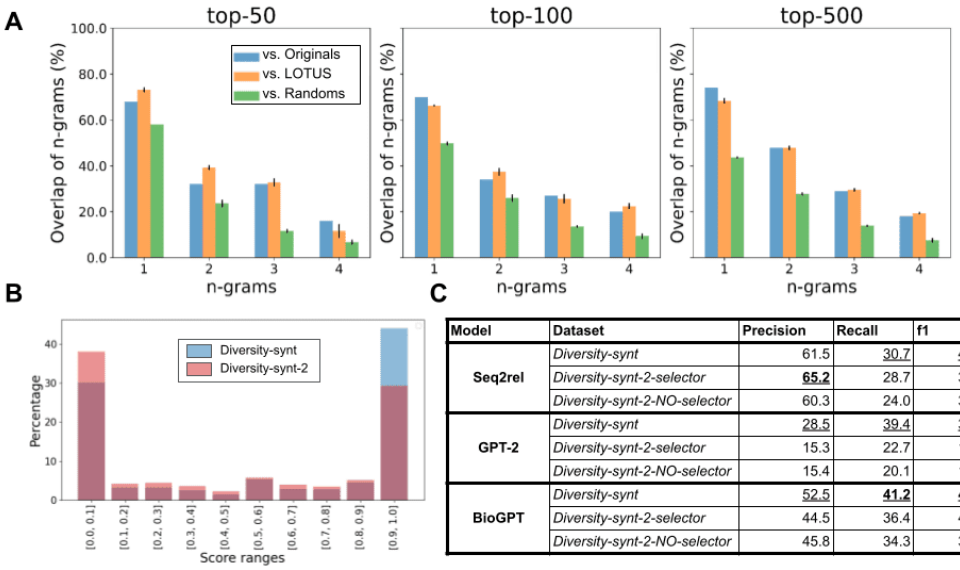
Figure A.8
An example of an abstract generation with a context hallucination. The original seed article is PMID 23758051. The generation was produced with $temp = 0.5$. It is a training example from the `Diversity-synt` dataset.

a dedicated decoding strategy. We fixed the temperature at $t = 2$ for all generations and used a top-k $= 500$ sampling strategy. With these decoding parameters, we intended to lower the quality of the generations by stimulating the "creativity" and increasing the frequency of instruction inconsistencies (Huang et al. 2023; Holtzman et al. 2020).

Panel B in Figure A.9 shows the distribution of the score obtained with the selector module between `Diversity-synt` and the newly created dataset `Diversity-synt-2`. Recall that the selector measures the proportion q of the relations, from the expected output labels, that have both their head and tail entities explicitly mentioned in the generated abstract. The observed shift clearly suggests more frequent inconsistencies between instructions and generated texts in `Diversity-synt-2`, where at least one member of a relation stated in the instructions is more frequently omitted.

Finally, we re-trained the 3 models (Seq2rel, GPT-2, and BioGPT) on two new training datasets from the new generations (with promoted hallucinations), and conducted an ablation study on the selector module. The dataset `Diversity-synt-2-selector` uses the implemented selector module to select the top-k $= 3$ generations per seed articles, while the selection was random for `Diversity-synt-2-NO-selector`. Seq2rel (encoder-decoder) performs robustly when trained on `Diversity-synt-2-selector`, while BioGPT and GPT-2 exhibit a more significant decrease in F1-score. However, all models show a decrease in performance when trained on `Diversity-synt-2-NO-selector`. Notably, Seq2rel and BioGPT models still perform better when trained on synthetic data with promoted hallucinations, than on the raw noisy data.

Together with these new results, our general observations suggest that generated abstracts exhibit typical lexical and syntactic features of the literature on natural products. The n-gram distribution of the synthetic data is more similar to the natural product literature than to random abstracts, and models trained on these data outperform models trained on the raw noisy data. The results on the newly generated datasets with promoted hallucinations show a decrease in performance across all trained models. This decrease, coupled with our analysis of the selector's score distribution, suggests that hallucinations, especially those related to the expression of the main findings (instruction inconsistencies), negatively impact model performance. The selector module can then alleviate this issue by excluding these undesirable generations.



Model	Dataset	Precision	Recall	f1
Seq2rel	*Diversity-synt*	61.5	30.7	40.1
	Diversity-synt-2-selector	**65.2**	28.7	39.9
	Diversity-synt-2-NO-selector	60.3	24.0	34.3
GPT-2	*Diversity-synt*	28.5	39.4	33.0
	Diversity-synt-2-selector	15.3	22.7	18.3
	Diversity-synt-2-NO-selector	15.4	20.1	17.4
BioGPT	*Diversity-synt*	52.5	**41.2**	**46.2**
	Diversity-synt-2-selector	44.5	36.4	40.0
	Diversity-synt-2-NO-selector	45.8	34.3	39.2

Figure A.9
A: Proportion of overlap in the top-50, top-100, and top-500 of the most frequent word n-grams (n = 1,2,3,4), from the `Diversity-synt` dataset, compared to 3 references: The original seed articles used for generation (*vs. Originals*), random articles sampled from LOTUS (*vs. LOTUS*), random articles from PubMed (*vs. Randoms*). The size of each reference set is equivalent to the size of the Original set ($n = 1{,}519$). For *vs. LOTUS* and *vs. Randoms*, the overlap values are averaged against 5 different random samples and the standard deviation is indicated. **B:** Histogram representing the distribution of the selector's score for all the generations between `Diversity-synt` and `Diversity-synt-2`. There are 15,190 generations per datasets (m = 10 generation per seed article). The `Diversity-synt-2` dataset is a new generated dataset using specific decoding parameters (temperature = 2 and top-k = 500) to promote hallucinations and instruction inconsistencies. **C:** Performance of fine-tuned models on new synthetic datasets with promoted hallucinations. Both were created using the same decoding strategy (temperature = 2 and top-k = 500), but the selector step was removed for `Diversity-synt-2-NO-selector`, resulting in a random selection of the generations. The respective size of the datasets are 3,045 and 3,562. `Diversity-synt-2-NO-selector` was sampled to match the size of the `Diversity-synt` dataset.

Appendix B. Evaluation Dataset

B.1 Dataset Curation Protocol

Biocurator: The dataset was curated by a single curator with a PhD in microbiology and prior experience in manual curation. A second annotator with a background in biology re-annotated the dataset to measure the inter-annotator agreement (IAA).

Article selection: Articles were selected using the proposed GME-sampler, by extracting the top-200 literature references which maximize the diversity of named entities. All selected articles have a PMID, an available abstract, a title, and are available online on PubMed. No filter was applied based on the journal or the publication date.

Objective: The curator targeted the relations between organisms (*head*) and their isolated natural products (*tail*) in the abstracts. Only organisms and chemicals that are

involved in NP relationships are extracted. For example, organisms on which the activity of a compound is tested (e.g., a pathogen like *Bacillus cereus*) are not annotated. The available LOTUS annotations were always used as a starting point.

Annotation of chemical entities: All chemical entities are categorized as either singular chemical (e.g., hispaglabridin A) or chemical classes (e.g., Isoflavanoids). The nature of these entities was cross-validated with the standard ChEBI Ontology when necessary. For singular chemicals, information about their chemical class is also extracted if it is mentioned in the article. Importantly, the label of the chemical entity is annotated as it is mentioned in the abstract. To align with the original LOTUS data, Wikidata and PubChem identifiers were assigned to chemicals and classes when available. In cases of ambiguity, the curator refers to the full-text (if available) to obtain more detailed information and assign the correct standardized entity. If the entity is not found in Wikidata, a dedicated identifier in the format "{pmid}CHEM{N}" is assigned instead, e.g., "11421752CHEM1".

Annotation of organism entities: Similarly to chemicals, the name of the organism is annotated exactly as it appears in the abstract. When only the genus is determined (e.g., Plakinastrella sp.), the genus name serves as the label.

Annotation of relations: The output labels only include relations explicitly mentioned in the abstract, while relations mentioned in the full-text are excluded. The relations are annotated based on their order of appearance in the abstract. If there are more than one organism, the relations of the first organism are annotated first, followed by the relations of the other organisms in order of appearance.

Export: The annotations are exported in a `JSON`-format as illustrated in Figure B.1 along with more statistics on the annotation.

B.2 Evaluation Dataset: Content Overview

An in-depth evaluation of the content of the curated dataset is provided in Figure B.2. The median number of relations, chemicals, and organisms, per curated abstracts are respectively 6, 5, and 1 (Panels A, B, C). Most of the studies included in the dataset focused on identifying natural products (up to max 22) from one specific organism. However, as illustrated in panels D and E, almost all chemicals and organisms only manifest once in the dataset, minimizing the overlap between the mentioned entities in each document. This is expected as a result of the diversity-sampling. Considering the applied stratification procedure, the distribution of the biological kingdoms (panel F) also shows a relatively balanced repartition.

The composition of the curated evaluation dataset, in terms of number of distinct entities and relationships, is also compared to 5 random sets of equivalent sizes in Table B.1. Firstly, 13 abstracts did not mention any relationships between organisms and chemicals in the curated dataset. Secondly, for the random sets, statistics were directly estimated from the LOTUS annotations. Then, they may represent an overestimate of the actual number of distinct entities, given that a manual curation could potentially eliminate some irrelevant annotations that are actually not mentioned in the abstracts. They should therefore be regarded as an approximate upper bound. Considering the

```
"19785389": {
    "PMID": "19785389",
    "ArticleTitle": "Further studies on the chemistry of the flustra alkaloids from the bryozoan Flustra foliacea.",
    "AbstractText": "Since 1980, over a dozen novel brominated alkaloids, named flustramines, have been isolated from
        Scandinavian and Canadian collections of the marine bryozoan Flustra foliacea. This paper describes the
        reisolation of the known compound dihydroflustramine C (1) and the isolation of 11 new flustramines (2-4, 6-13), including
        two dimers (12, 13) that may be isolation artifacts. …",
    "id": "Q51648886",
    "doi": "10.1021/NP900282J",
    "chemicals": [
      {
        "id": "Q5863357",
        "type": "class",
        "label": "flustramine"
      },
      {
        "id": "Q82888434",
        "pubchem_id": "185307",
        "type": "chemical",
        "label": "dihydroflustramine C",
        "class": "flustramines"
      }
    ],
    "organisms": [
      {
        "id": "Q2399981",
        "label": "Flustra foliacea"
      }
    ],
    "relations": [
      [
        "Q2399981",
        "Q5863357"
      ],
      [
        "Q2399981",
        "Q82888434"
      ]
    ]
  }
```


Chemical names in the abstract	% of abstracts
Chemical name	48.7%
Class names	10.7%
Chemical and class names	37.1%
No name	3.5%

Organism names in the abstract	% of abstracts
Species name	74%
Genus name	16%
Genus and species names	4.5%
No name	5.5%

Figure B.1
An example of a curated literature reference in the evaluation dataset with supplementary statistics.

Table B.1
Statistics of the number of organisms, chemicals, and relations in the top-200 abstracts selected and curated in the evaluation set, compared to 200 randomly selected items (statistics averaged over 5 random seeds). For the evaluation set, the number of annotated distinct chemical compounds and chemical classes are respectively indicated between parentheses. In the curated evaluation set, 13 references had no relation directly expressed in the abstract.

	# Organisms	# Chemicals	# Relations	# References
eval-set (top-200 diversity)	275	1,197 (1,092 / 105)	1,488 (1,297 / 191)	200 (187*)
Random (200 articles)	238	610	699	200

last points, the proposed strategy for selecting the evaluation set has significantly improved the diversity.

B.3 Inter-annotator Agreement

To assess the quality of the annotations in the evaluation dataset, we computed the IAA for the extracted relationships, following the same method as in Li et al. (2016). We use the Jaccard Index to measure the IAA, considering the union of all the extracted relations with a second annotator, who followed the same guidelines. A disagreement between the annotators occurred when there was a mismatch in the label or type of the chemical ("*chemical*" or "*class*"), or, in the label of the organism.

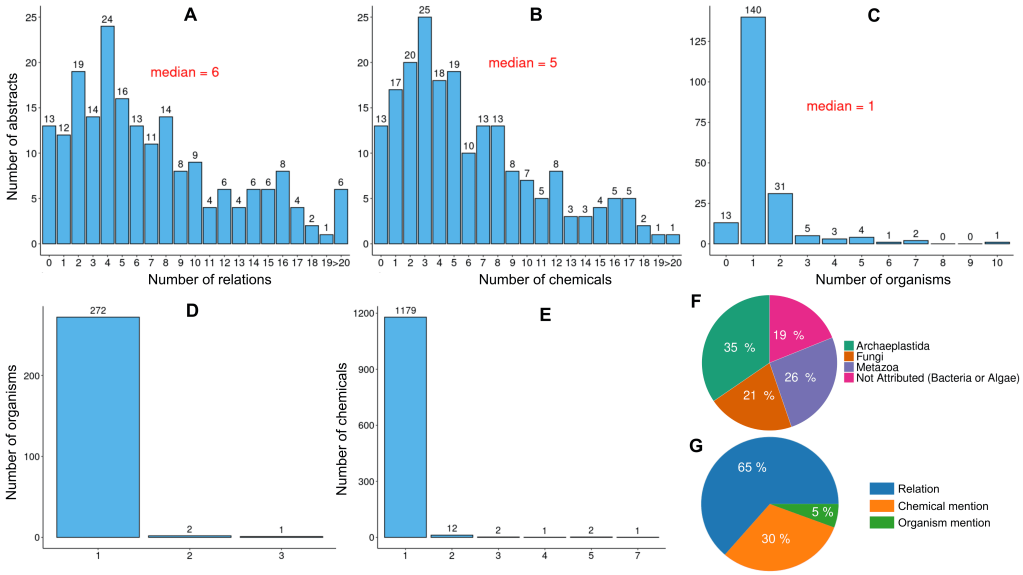


Figure B.2
Panels **A**, **B**, and **C** respectively represent the distribution of the number of relations, chemicals, and organisms per annotated abstracts (200). Panel **D** represents the distribution of the frequency of mention of the 275 distinct organisms in the corpus, and similarly for chemicals in panel **E**. Panel **F** shows the distribution of the biological kingdoms of the annotated organisms in the curated dataset. Panel **G** is the repartition of the sources of disagreement between annotations gathered from the annotators. The 179 disagreements are qualified in 3 categories: *Organism mention* disagreement occurs when the 2 annotators disagree on the label of an organism entity; *Chemical mention* disagreement occurs when either the name or type of the extracted chemical mentioned differed; *Relation* disagreement arises when annotators disagree on the status of the relationship between a chemical entity and an organism in the text.

The observed IAA score is 88.5%. Out of the 1,569 annotations provided by the two annotators, 179 were subject to disagreements. An analysis of the disagreements is provided in panel G of Figure B.2. For example, in PMID 16595963, the first annotator annotated the compound 4 as "GS-4", while the second annotator used the later identification "(4R,4aS,9aR)-1,9a-dihydronidulalin A" (*Chemical mention* disagreement). In another example, in PMID 32193929, the ambiguous links between "oudemansins", "oudemansinols", "polyketides", and "*Favolaschia calocera*" were the subject of a disagreement between the annotators on the status of the relationships (*Relation* disagreement). Overall, while the identification of organisms involved in a relation from the text was almost always in agreement, the main sources of disagreements concern the status of relationships and the identifications (the extracted labels) of the chemicals (or classes).

Appendix C. Supplementary Materials

C.1 Chemical Length Thresholding

To determine a reasonable threshold for filtering chemical labels with excessive length, we conducted a comparative analysis of the distribution of label lengths in LOTUS

(derived from Wikidata) versus their corresponding IUPAC names (See Figure C.1). While the respective median and mean values clearly suggest that most of the available chemicals are identified with common names (i.e., shorter), the long right-tail of labels exhibit a length comparable to IUPAC names. These longer labels are often too lengthy to be practical for use in training examples for the targeted RE task. By estimating the limit when 90% of the chemical labels in LOTUS are at least as long as their corresponding IUPAC name, we estimated that a threshold of 60 characters effectively filters out excessively long labels.

C.2 Mismatches Between Standardized Labels and Original Abstracts

The 7,901 available abstracts from the literature references in the *Extended* dataset were extracted using the NCBI E-utilities `efectch` service. All the organism labels available on Wikidata were directly matched on the abstracts. Using the PubChem exchange service, all the synonyms (direct synonyms of the molecule and synonyms of its stereoisomers) were extracted when a PubChem ID was available. In total, 653,749 synonyms were extracted. A chemical entity was considered as mentioned in the abstract when there is an exact match of its name or one of its synonym in the abstract. Some chemicals, however, may also only be implicitly mentioned in an abstract. Indeed, the isolation of multiple derivatives, such as Atroviridin A, B, and C, is typical reported as "Atroviridins A-C". Then, Atroviridin B would not be explicitly mentioned and has to be infered. All chemicals which could be part of such expressions were identified using a set of regular expressions and were treated separately to not wrongly inflate

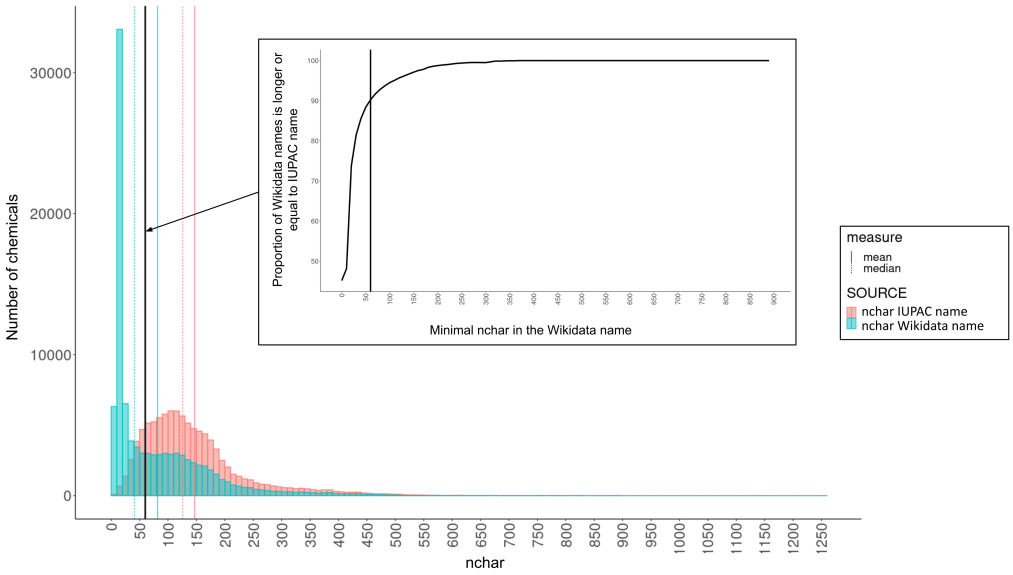


Figure C.1
Distribution of the number of characters in chemical names between Wikidata labels and their corresponding IUPAC name. The integrated graph indicates the proportion of chemicals for which the length of the Wikidata label is longer or equal to the IUPAC name, by increasing the number of characters in the Wikidata label. The black vertical line corresponds to the chosen threshold at 60 characters.

the proportion of chemicals not mentioned in the abstracts. Nonetheless, it is worth mentioning that a non-negligeable part of these *multiple* chemical entities are simply not mentioned in the abstract, either explicitly or implicitly. For instance, see the original mentions of *malyngamide A*[21] in PMID 10924193, 11076568, and 21341718.

C.3 Raw and Synthetic Datasets Overview

Table C.1
Impact of the pre-processing of the number of organisms, chemicals, and relations.

	# Organisms	# Chemicals	# Relations	# References
Original dataset	36,803	220,783	533,347	88,810
Pre-processed dataset	14,890	56,310	102,528	32,616

Table C.2
Maximal number N of literature items per kingdoms along with the value and the rank of the maximal reached entropies on organisms $H_S(O)$ and chemicals $H_S(C)$.

Kingdom	**N**	**max** $H_S(O)$ **(rank)**	**max** $H_S(C)$ **(rank)**
Archaeplastida	19,491	8.73 (10,512)	9.81 (10,713)
Fungi	5,023	7.18 (2,519)	9.33 (5,023)
Metazoa	1,920	6.72 (1,304)	8.33 (1,920)
Not Attributed (Bacteria or Algae)	6,666	6.96 (2,503)	8.90 (6,666)

Table C.3
Percentage of the maximal (observed) entropies $H_S(O)$ and $H_S(C)$ at different steps: 250, 500, 1,000, and 2,000 top-ranked articles.

	Organisms (% of Max Entropy)				**Chemicals (% of Max Entropy)**			
Kingdom	**n = 250**	**n = 500**	**n = 1,000**	**n = 2,000**	**n = 250**	**n = 500**	**n = 1,000**	**n = 2,000**
Archaeplastida	75.5	80.5	86	91.7	76.9	83.7	89.6	94.3
Fungi	80.7	89	96.6	99.8	83.7	88.1	92.3	95.1
Metazoa	84.4	93	99.5	96.1	90	94.6	97.4	100
Not Attributed (Bacteria or Algae)	82.3	90.7	96.7	99.9	85.1	89.6	93.7	97.2

C.4 Evaluation of Keyword Extraction on the SemEVAL2017 Dataset

The SemEVAL2017 (Augenstein et al. 2017) evaluation dataset consists of 100 paragraphs, extracted from scientific publications in various domains, with on average 17.23 annotated keyphrases. While 3 sub-tasks are proposed in this challenge (classification and semantic relations), we only focused on the mention-level keyphrases identification. To consider similar settings as used for synthetic abstract generation, we evaluated the precision in the top-10 extracted keywords. The comparison is done by exact-match and results are presented in Table C.5. Vicuna-13B largely outperforms the KeyBERT

21 `https://www.wikidata.org/wiki/Q27135775`.

Table C.4
Statistics on the content of the created datasets. For `Diversity-raw`, the top-50 articles per biological kingdoms (with an available abstract) were reserved for the evaluation set. The train/valid sets are composed of the remaining items split in 90:10. A similar split was performed on the initial 5 random samples to obtain the train/valid datasets of equivalent sizes, referred as the `Random-raw` datasets. Their count statistics are averaged over the 5 seeds. `Extended-raw` is the fusion of the `Diversity-raw` plus the 5 `Random-raw` datasets. `Full` is a dataset containing all available examples from the LOTUS snapshot, except the 200 used in the evaluation set. For synthetic datasets, the number of relations, as well as the number of distinct chemicals, is split between chemical entities and chemical classes.

Dataset	Part.	# Relations (w. chem / w. class)	# Organisms	# Chemical entities (chem. / class.)	# References
`Diversity-raw`	train	12,666	2,644	10,311	1,519*
	valid	1,425	301	1,211	168*
`Random-raw`	train	5,102	1,434	4,286	1,531*
	valid	657	220	584	189*
`Extended-raw`	train	27,952	5,642	21,028	7,111*
	valid	3,355	932	2,741	790*
`Full`	train	90,326	13,208	51,658	28,286
	valid	1,533	484	1,288	430
`Diversity-synt`	train	11,547 (10,764 / 783)	2,154	(9,108 / 61)	3,562
	valid	1,197 (1,096 / 101)	220	(998 / 37)	389
`Random-synt`	train	4,825 (4,474 / 351)	1,267	(3,854 / 53)	3,798
	valid	609 (561 / 47)	190	(507 / 22)	460
`Extended-synt`	train	28,614 (26,373 / 2,242)	5,258	(20,404 / 69)	23,985
	valid	1,444 (1,332 / 112)	432	(1,122 / 37)	1,254

Table C.5
Comparison of the performance of the prompted Vicuna-13B LLM and KeyBERT for keywords/keyphrases extraction on the *SemEVAL2017* test set. The evaluation was only done on the top-10 extracted keywords for both methods, to use the same configuration as in the experiments.

	TP in Top-10	FP in Top-10	Precision
KeyBERT (all-MiniLM-L6-v2)	96	904	9.6
Vicuna-13B	321	669	**32.424**

(Grootendorst 2020) baseline and shows more than acceptable performance in zero-shot settings. KeyBERT was used with standard parameters: `keyphrase_ngram_range: (1,2), stop_words: None, use_mmr: True, diversity: 0.7` and BERT model `all-MiniLM-L6-v2` for base embeddings.

C.5 Tokenized Length of Abstracts

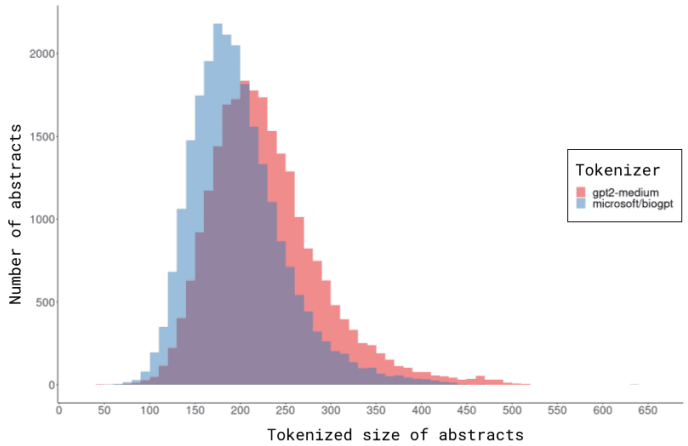


Figure C.2
Differences in size of the tokenized abstracts from the *Extended-raw* dataset (23,985 abstracts) between GPT-2 and BioGPT tokenizers. The dedicated tokenizer of BioGPT allows for a more efficient tokenization of the abstracts.

Acknowledgments

The authors are thankful to Vincent Mutel, Joël Dumoulin, Joel Rossier, and Colombine Verzat for their help during the project. We are grateful to Olena Hrynenko for proofreading the mathematical formulations. We are also grateful to the authors behind LOTUS, BioGPT, and Seq2rel for sharing their data or code.

Funding

This work was supported by the IDIAP Research Institute and has been done in collaboration with the company Inflamalps SA and is supported by the Ark Foundation. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 965397. The funding bodies played no role in the design of the study, research, writing, and publication of the article.

References

Aggarwal, Karan, Henry Jin, and Aitzaz Ahmad. 2023. ECG-QALM: Entity-controlled synthetic text generation using contextual Q&A for NER. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5649–5660. https://doi.org/10.18653/v1/2023.findings-acl.349

Aghajanyan, Armen, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328. https://doi.org/10.18653/v1/2021.acl-long.568

Ahmed, Mahtab, Jumayel Islam, Muhammad Rifayat Samee, and Robert E. Mercer. 2019. Identifying protein-protein interaction using tree LSTM and structured attention. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 224–231. https://doi.org/10.1109/ICOSC.2019.8665584

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631. https://doi.org/10.1145/3292500.3330701

Anaby-Tavor, Ateret, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? Deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390. https://doi.org/10.1609/aaai.v34i05.6233

Augenstein, Isabelle, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555. https://doi.org/10.18653/v1/S17-2091

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Bonifacio, Luiz, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Data augmentation for information retrieval using large language models. *ArXiv:2202.05144*. https://doi.org/10.1145/3477495.3531863

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Cao, Boxi, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? Revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874. https://

doi.org/10.18653/v1/2021.acl-long.146

Chen, Jiaao, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211. https://doi.org/10.1162/tacl_a_00542

Chen, Maximillian, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022a. Weakly supervised data augmentation through prompting for dialogue understanding. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Chen, Yanda, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022b. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730. https://doi.org/10.18653/v1/2022.acl-long.53

Chia, Yew Ken, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2024. InstructEval: Towards holistic evaluation of instruction-tuned large language models. In *Proceedings of the First Edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 35–64.

Chiang, Wei Lin, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. See https://vicuna.lmsys.org (accessed 14 April 2023). 2(3):6.

Chung, John Joon Young, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593. https://doi.org/10.18653/v1/2023.acl-long.34

Dai, Zhuyun, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Dettmers, Tim, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Eberts, Markus and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660. https://doi.org/10.18653/v1/2021.eacl-main.319

Fan, Yang, Fei Tian, Tao Qin, Jiang Bian, and Tie-Yan Liu. 2017. Learning what data to learn. *ArXiv:1702.08635*.

Feng, Jun, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):5779–5786. https://doi.org/10.1609/aaai.v32i1.12063

Feng, Steven Y., Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84

Galvao, Roberto Kawakami Harrop, Mário César Ugulino Araujo, Gledson Emídio José, Marcio José Coelho Pontes, Edvan Cirino Silva, and Teresa Cristina Bezerra Saldanha. 2005. A method for calibration and validation subset partitioning. *Talanta*, 67(4):736–740. https://doi.org/10.1016/j.talanta.2005.03.025, PubMed: 18970233

Gao, Jiahui, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*.

Gerner, Martin, Goran Nenadic, and Casey M. Bergman. 2010. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85.

https://doi.org/10.1186/1471-2105-11-85, PubMed: 20149233
Giorgi, John, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25. https://doi.org/10.18653/v1/2022.bionlp-1.2
Gong, Zhiqiang, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *IEEE Access*, 7:64323–64350. https://doi.org/10.1109/ACCESS.2019.2917620
Grootendorst, Maarten. 2020. KeyBERT: Minimal keyword extraction with BERT.
Hartvigsen, Thomas, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326. https://doi.org/10.18653/v1/2022.acl-long.234
He, Xuanli, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842. https://doi.org/10.1162/tacl_a_00492
Hill, M. O. 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432. https://doi.org/10.2307/1934352
Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
Hou, Yutai, Yingce Xia, Lijun Wu, Shufang Xie, Yang Fan, Jinhua Zhu, Tao Qin, and Tie-Yan Liu. 2022. Discovering drug–target interaction knowledge from biomedical literature. *Bioinformatics*, 38(22):5100–5107. https://doi.org/10.1093/bioinformatics/btac648, PubMed: 36205562
Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
Hu, Xuming, Aiwei Liu, Zeqi Tan, Xin Zhang, Chenwei Zhang, Irwin King, and Philip S. Yu. 2023. GDA: Generative data augmentation techniques for relation extraction tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10221–10234. https://doi.org/10.18653/v1/2023.findings-acl.649
Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.
Huguet Cabot, Pere Lluís, and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381. https://doi.org/10.18653/v1/2021.findings-emnlp.204
Iinuma, Naoki, Makoto Miwa, and Yutaka Sasaki. 2022. Improving supervised drug-protein relation extraction with distantly supervised models. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 161–170. https://doi.org/10.18653/v1/2022.bionlp-1.16
Jain, Abhinav, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3561–3562. https://doi.org/10.1145/3394486.3406477
Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv:2401.04088*.
Jimenez Gutierrez, Bernal, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? Think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

https://doi.org/10.18653/v1/2022.findings-emnlp.329

Joseph, V. Roshan and Akhil Vakayil. 2022. SPlit: An optimal method for data splitting. *Technometrics*, 64(2):166–176. https://doi.org/10.1080/00401706.2021.1921037

Josifoski, Martin, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574. https://doi.org/10.18653/v1/2023.emnlp-main.96

Jost, Lou. 2006. Entropy and diversity. *Oikos*, 113(2):363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kambar, Mina Esmail Zadeh Nojoo, Armin Esmaeilzadeh, and Kazem Taghva. 2022. Chemical-gene relation extraction with graph neural networks and BERT encoder. In *Proceedings of the ICR'22 International Conference on Innovations in Computing Research*, pages 166–179. https://doi.org/10.1007/978-3-031-14054-9_17

Kambar, Mina Esmail Zadeh Nojoo, Armin Esmaeilzadeh, and Maryam Heidari. 2022. A survey on deep learning techniques for joint named entities and relation extraction. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 218–224. https://doi.org/10.1109/AIIoT54504.2022.9817231

Kennard, R. W. and L. A. Stone. 1969. Computer aided design of experiments. *Technometrics*, 11(1):137–148. https://doi.org/10.1080/00401706.1969.10490666

Kim, Hyuhng Joon, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *CoRR*, abs/2206.08082.

Kim, Hyun Woo, Mingxun Wang, Christopher A. Leber, Louis-Félix Nothias, Raphael Reher, Kyo Bin Kang, Justin J. J. Van Der Hooft, Pieter C. Dorrestein, William H. Gerwick, and Garrison W. Cottrell. 2021. NPClassifier: A deep neural network-based structural classification tool for natural products. *Journal of Natural Products*, 84(11):2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399, PubMed: 34662515

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Krallinger, Martin, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1):S2. https://doi.org/10.1186/1758-2946-7-S1-S2, PubMed: 25810773

Kumar, Varun, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Labrak, Yanis, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. *ArXiv:2402.10373*.

Leinster, Tom. 2021. *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press. https://doi.org/10.1017/9781108963558

Li, Fei, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1):198. https://doi.org/10.1186/s12859-017-1609-9, PubMed: 28359255

Li, Jiao, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database: The Journal of*

Biological Databases and Curation, 2016:baw068. https://doi.org/10.1093/database/baw068, PubMed: 27161011

Li, Ming, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. *ArXiv:2308.12032*. https://doi.org/10.48550/arXiv.2308.12032

Li, Xu, Jing Ren, Wen Zhang, Zhiming Zhang, Jinchao Yu, Jiawei Wu, He Sun, Shuiping Zhou, Kaijing Yan, Xijun Yan, and Wenjia Wang. 2022. LTM-TCM: A comprehensive database for the linking of traditional Chinese medicine with modern medicine at molecular and phenotypic levels. *Pharmacological Research*, 178:106185. https://doi.org/10.1016/j.phrs.2022.106185, PubMed: 35306140

Liang, Weixin, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8):669–677. https://doi.org/10.1038/s42256-022-00516-1

Liu, Jiachang, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114. https://doi.org/10.18653/v1/2022.deelio-1.10

Luo, Ling, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022a. BioRED: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282. https://doi.org/10.1093/bib/bbac282, PubMed: 35849818

Luo, Renqian, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409. https://doi.org/10.1093/bib/bbac409, PubMed: 36156661

Mallen, Alex, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

Mangrulkar, Sourab, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft

Mazumder, Mark, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William A. Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Daniel Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Y. Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2023. DataPerf: Benchmarks for data-centric AI development. In *Advances in Neural Information Systems*, volume 36, pages 5320–5347.

Meng, Yu, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 462–477.

Meng, Yu, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24457–24477.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011. https://doi.org/10.3115/1690219.1690287

Newman, Mark E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351. https://doi.org/10.1080/00107510500052444

Northcutt, Curtis, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411. https://doi.org/10.1613/jair.1.12125

Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, pages 1373–1411.

Paolini, Giovanni, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Papanikolaou, Yannis and Andrea Pierleoni. 2020. DARE: Data augmented relation extraction with GPT-2. *ArXiv: 2004.13845*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics - ACL '02*, page 311. https://doi.org/10.3115/1073083.1073135

Pellicer, Lucas Francisco Amaral Orosco, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803. https://doi.org/10.1016/j.asoc.2022.109803

Prusa, Joseph, Taghi M. Khoshgoftaar, and Naeem Seliya. 2015. The effect of dataset size on training tweet sentiment classifiers. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 96–102. https://doi.org/10.1109/ICMLA.2015.22

Pu, Dongqi and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. *ArXiv:2306.07799*. https://doi.org/10.18653/v1/2023.acl-srw.1

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Rutz, Adriano, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G. Graham, Ralf Stephan, Roderic Page, Jirí Vondrášek, Christoph Steinbeck, Guido F. Pauli, Jean-Luc Wolfender, Jonathan Bisson, and Pierre-Marie Allard. 2022. The LOTUS initiative for open knowledge management in natural products research. *eLife*, 11:e70780. https://doi.org/10.7554/eLife.70780, PubMed: 35616633

Ríos-Toledo, Germán, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez. 2022. Detection of changes in literary writing style using N-grams as style markers and supervised machine learning. *PLOS ONE*, 17(7):e0267590. https://doi.org/10.1371/journal.pone.0267590, PubMed: 35857768

Salhofer, Eileen, Xing Lan Liu, and Roman Kern. 2022. Impact of training instance selection on domain-specific entity extraction using BERT. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 83–88. https://doi.org/10.18653/v1/2022.naacl-srw.11

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15. https://doi.org/10.1145/3411764.3445518

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon,

Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Schick, Timo and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951. https://doi.org/10.18653/v1/2021.emnlp-main.555

Shahab, Elham. 2017. A short survey of biomedical relation extraction techniques. *ArXiv:1707.05850*.

Shang, Jingbo, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064. https://doi.org/10.18653/v1/D18-1230

Shinbo, Y., Y. Nakamura, Md Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya. 2006. KNApSAcK: A comprehensive species-metabolite relationship database. *Plant Metabolomics*, pages 165–181. https://doi.org/10.1007/3-540-29782-0_13

Si, Chenglei, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Sidorov, Grigori, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860. https://doi.org/10.1016/j.eswa.2013.08.015

Smirnova, Alisa and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys*, 51(5):106:1–106:35. https://doi.org/10.1145/3241741

Smith, Ryan, Jason A. Fries, Braden Hancock, and Stephen H. Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM / IMS Journal of Data Science*, 1(2):1–30. https://doi.org/10.1145/3617130

Sorokina, Maria, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik, and Christoph Steinbeck. 2021. COCONUT online: Collection of open natural products database. *Journal of Cheminformatics*, 13(1):1–13. https://doi.org/10.1186/s13321-020-00478-9, PubMed: 33423696

Stefanini, Irene, Silvia Carlin, Noemi Tocci, Davide Albanese, Claudio Donati, Pietro Franceschi, Michele Paris, Alberto Zenato, Silvano Tempesta, Alberto Bronzato, Urska Vrhovsek, Fulvio Mattivi, and Duccio Cavalieri. 2017. Core microbiota and metabolome of Vitis vinifera L. cv. corvina grapes and musts. *Frontiers in Microbiology*, 8:Art. 457. https://doi.org/10.3389/fmicb.2017.00457, PubMed: 28377754

Su, Junhao, Ye Wu, Hing-Fung Ting, Tak-Wah Lam, and Ruibang Luo. 2021. RENET2: High-performance full-text gene–disease relation extraction with iterative training data expansion. *NAR Genomics and Bioinformatics*, 3(3):lqab062. https://doi.org/10.1093/nargab/lqab062, PubMed: 34235433

Su, Peng, Gang Li, Cathy Wu, and K. Vijay-Shanker. 2019. Using distant supervision to augment manually annotated data for relation extraction. *PLOS ONE*, 14(7):e0216913. https://doi.org/10.1371/journal.pone.0216913, PubMed: 31361753

Swainston, Neil, Kieran Smallbone, Hooman Hefzi, Paul D. Dobson, Judy Brewer, Michael Hanscho, Daniel C. Zielinski, Kok Siong Ang, Natalie J. Gardiner, Jahir M. Gutierrez, Sarantos Kyriakopoulos, Meiyappan Lakshmanan, Shangzhong Li, Joanne K. Liu, Veronica S. Martínez, Camila A. Orellana, Lake-Ee Quek, Alex Thomas, Juergen Zanghellini, Nicole Borth, Dong-Yup Lee, Lars K. Nielsen, Douglas B. Kell, Nathan E. Lewis, and Pedro Mendes. 2016. Recon 2.2: From reconstruction to model of human metabolism. *Metabolomics*, 12(7):109. https://doi.org/10.1007/s11306-016-1051-4, PubMed: 27358602

Tang, Ruixiang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? *ArXiv:2303.04360*.

Thiele, Ines, Neil Swainston, Ronan M. T. Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K. Aurich, Hulda

Haraldsdottir, Monica L. Mo, Ottar Rolfsson, Miranda D. Stobbe, Stefan G. Thorleifsson, Rasmus Agren, Christian Bölling, Sergio Bordel, Arvind K. Chavali, Paul Dobson, Warwick B. Dunn, Lukas Endler, David Hala, Michael Hucka, Duncan Hull, Daniel Jameson, Neema Jamshidi, Jon J. Jonsson, Nick Juty, Intawat Nookaew, Nicolas Le Novére, Naglis Malys, Alexander Mazein, Jason A. Papin, Nathan D. Price, Evgeni Selkov, Martin I. Sigurdsson, Evangelos Simeonidis, Nikolaus Sonnenschein, Kieran Smallbone, Anatoly Sorokin, Johannes H. G. M. van Beek, Dieter Weichart, Igor Goryanin, Jens Nielsen, Hans V. Westerhoff, Douglas B. Kell, Pedro Mendes, and Bernhard O. Palsson. 2013. A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, 31(5):419–425. https://doi.org/10.1038/nbt.2488, PubMed: 23455439

Thompson, Lilian U., Beatrice A. Boucher, Zhen Liu, Michelle Cotterchio, and Nancy Kreiger. 2006. Phytoestrogen content of foods consumed in Canada, including isoflavones, lignans, and coumestan. *Nutrition and Cancer*, 54(2):184–201. https://doi.org/10.1207/s15327914nc5402_5, PubMed: 16898863

Veselovsky, Veniamin, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *ArXiv:2305.15041*.

Wang, Difeng, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721. https://doi.org/10.18653/v1/2020.emnlp-main.303

Wang, Zirui, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *ArXiv:2109.09193*.

Wei, Chih-Hsuan, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593. https://doi.org/10.1093/nar/gkz389, PubMed: 31114887

Wu, Chaoyi, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045. https://doi.org/10.1093/jamia/ocae045, PubMed: 38613821

Wu, Zhiyong, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436. https://doi.org/10.18653/v1/2023.acl-long.79

Wysocki, Oskar, Zili Zhou, Paul O'Regan, Deborah Ferreira, Magdalena Wysocka, Dónal Landers, and André Freitas. 2023. Transformers and the representation of biomedical background knowledge. *Computational Linguistics*, 49(1):73–115. https://doi.org/10.1162/coli_a_00462

Xu, Benfeng, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207. https://doi.org/10.18653/v1/2023.acl-long.455

Xu, Xin, Xiang Chen, Ningyu Zhang, Xin Xie, Xi Chen, and Huajun Chen. 2022. Towards realistic low-resource relation extraction: A benchmark with empirical baseline study. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 413–427. https://doi.org/10.18653/v1/2022.findings-emnlp.29

Xu, Yun and Royston Goodacre. 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3):249–262. https://doi.org/10.1007/s41664-018-0068-2, PubMed: 30842888

Yang, Yiben, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025. https://doi.org/10.18653/v1/2020.findings-emnlp.90

Ye, Jiacheng, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669. https://doi.org/10.18653/v1/2022.emnlp-main.801

Yoo, Kang Min, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239. https://doi.org/10.18653/v1/2021.findings-emnlp.192

Yoon, Wonjin, Sean Yi, Richard Jackson, Hyunjae Kim, Sunkyu Kim, and Jaewoo Kang. 2023. Biomedical relation extraction with knowledge base–refined weak supervision. *Database*, 2023:baad054. https://doi.org/10.1093/database/baad054, PubMed: 37551911

Yu, Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*.

Zeng, Xiangrong, Shizhu He, Daojian Zeng, Kang Liu, Shengping Liu, and Jun Zhao. 2019. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 367–377. https://doi.org/10.18653/v1/D19-1035

Zeng, Xiangrong, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514. https://doi.org/10.18653/v1/P18-1047, PubMed: 30207992

Zha, Daochen, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *ArXiv:2303.10158*. https://doi.org/10.5772/intechopen.111542

Zhang, Ranran Haoran, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. 2020. Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 236–246. https://doi.org/10.18653/v1/2020.findings-emnlp.23

Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *ArXiv:2308.10792*.

Zhang, Tianlin, Jiaxu Leng, and Ying Liu. 2020. Deep learning for drug-drug interaction extraction from the literature: A review. *Briefings in Bioinformatics*, 21(5):1609–1627. https://doi.org/10.1093/bib/bbz087, PubMed: 31686105

Zhao, Sendong, Chang Su, Chang Su, Zhiyong Lu, Zhiyong Lu, Zhiyong Lu, and Fei Wang. 2020. Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, 22(3):1–19. https://doi.org/10.1093/bib/bbaa057, PubMed: 32422651

Zhao, Xiaoyan, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2023. A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers. *CoRR*, abs/2306.02051.

Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706.

Zhou, Chunting, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 55006–55021.